



PATENT
Attorney Docket No.: 16869N-104900US
Client Ref. No.: NT1432US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

KATSUYA TANAKA et al.

Application No.: 10/770,723

Filed: February 2, 2004

For: STORAGE DEVICE AND
CONTROLLING METHOD
THEREOF

Customer No.: 20350

Examiner: Unassigned

Technology Center/Art Unit: 2655

Confirmation No.: 2283

**PETITION TO MAKE SPECIAL FOR
NEW APPLICATION UNDER M.P.E.P.
§ 708.02, VIII & 37 C.F.R. § 1.102(d)**

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is a petition to make special the above-identified application under MPEP § 708.02, VIII & 37 C.F.R. § 1.102(d). The application has not received any examination by an Examiner.

(a) The Commissioner is authorized to charge the petition fee of \$130 under 37 C.F.R. § 1.17(i) and any other fees associated with this paper to Deposit Account 20-1430.

(b) All the claims are believed to be directed to a single invention. If the Office determines that all the claims presented are not obviously directed to a single invention, then Applicants will make an election without traverse as a prerequisite to the grant of special status.

05/02/2005 RFEKADU1 00000004 201430 10770723

01 FC:1464 130.00 DA

(c) Pre-examination searches were made of U.S. issued patents, including a classification search and a key word search. The classification search was conducted on or around April 1, 2005 covering Class 369 (subclasses 47.15 and 47.36), Class 710 (subclass 316), and Class 711 (subclasses 101, 111, 112, 113, 114, 154, and 170), by a professional search firm, Lacasse & Associates, LLC. The key word search was performed on the USPTO full-text database including published U.S. patent applications. The inventors further provided four references considered most closely related to the subject matter of the present application (see references #7-10 below), which were cited in the Information Disclosure Statement filed with the application on February 2, 2004.

(d) The following references, copies of which are attached herewith, are deemed most closely related to the subject matter encompassed by the claims:

- (1) U.S. Patent No. 6,295,587 B1;
- (2) U.S. Patent No. 6,587,919 B2;
- (3) U.S. Patent No. 6,640,281 B2;
- (4) U.S. Patent No. 6,701,411 B2;
- (5) U.S. Patent Publication No. 2003/0191891 A1;
- (6) U.S. Patent Publication No. 2005/0027919 A1;
- (7) Japanese Patent Publication No. JP 2003-85117; ✓
- (8) Japanese Patent Publication No. JP 2000-222339; ✓
- (9) Japanese Patent Publication No. JP 2003-303055; and ✓
- (10) Qlogic Corp., "Full Duplex and Fibre Channel, Network Storage Group Host Products Technology Brief."

(e) Set forth below is a detailed discussion of references which points out with particularity how the claimed subject matter is distinguishable over the references.

A. Claimed Embodiments of the Present Invention

The claimed embodiments relate to controlling data transfer of a storage device.

Independent claim 1 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. A destination drive I/O port to which a frame is to be forwarded is determined, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives.

Independent claim 4 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. A path which a frame passes to be transferred between the switch and one of the disk drives is determined, according to the type of a command included in an exchange between the disk adapter and the one of the disk drives.

Independent claim 6 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. The disk adapter determines destination information within a frame to be transferred from the disk adapter to one of the disk drives, according the type of a command included in an exchange between the disk adapter and the one of the disk drives. The switch selects one of port to port connection paths between a port to which the disk adapter is connected and ports to which the disk drives constituting the disk array are connected to switch each frame inputted to the switch, according to the destination information within the frame.

Independent claim 7 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. A destination drive port to which a frame is to be forwarded is determined, depending on whether the type of a command included in an

exchange that is transferred between the disk adapter and one of the disk drives is a data read command or a data write command. The exchange for reading data and the exchange for writing data are executed in parallel.

Independent claim 8 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; and a disk array comprising disk drives, each being equipped with a plurality of I/O ports. The disk adapter and the disk array are connected via a switch. A path which a frame passes between the switch and one of the disk drives is determined, depending on whether the type of a command included in an exchange between the disk adapter and the one of the disk drives is a data read command or a data write command.

Independent claim 9 recites a disk device comprising a disk controller comprising a channel adapter, a cache memory, and a disk adapter; a plurality of disk drives, each being equipped with a plurality of I/O ports; and a switch connecting the disk controller and the plurality of disk drives. A destination drive port to which a frame is to be forwarded is determined, depending on whether the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives is a data read command or a data write command. The exchange for reading data and the exchange for writing data are executed in parallel.

One of the benefits that may be derived is that a disk device having a back-end network that enables full duplex data transfer by simple control techniques can be realized, and that increased disk device throughput is achieved.

B. Discussion of the References

None of the following references disclose determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9. For instance, claim 1 recites a destination drive I/O port to which a frame is to be forwarded is determined, according to the type of a command included in an exchange that is transferred between the disk adapter

and one of the disk drives. Claim 4 recites a path which a frame passes to be transferred between the switch and one of the disk drives is determined, according to the type of a command included in an exchange between the disk adapter and the one of the disk drives. Claim 6 recites that the disk adapter determines destination information within a frame to be transferred from the disk adapter to one of the disk drives, according the type of a command included in an exchange between the disk adapter and the one of the disk drives. Claim 7 recites a destination drive port to which a frame is to be forwarded is determined, depending on whether the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives is a data read command or a data write command. Claim 8 recites a path which a frame passes between the switch and one of the disk drives is determined, depending on whether the type of a command included in an exchange between the disk adapter and the one of the disk drives is a data read command or a data write command. Claim 9 recites a destination drive port to which a frame is to be forwarded is determined, depending on whether the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives is a data read command or a data write command.

1. U.S. Patent No. 6,295,587 B1

The patent to Martin (6,295,587 B1), assigned to EMC Corp., provides for a Method and Apparatus for Multiple Disk Drive Access in a Multi-Processor/Multi-Disk Drive System. Disclosed is disk device 14, drive adapters 18, and adapter ports (AP) 20 and drive ports (DP) 24 of switch 22. AP 20 of switch 22 is connected to drive adapters 18 and DP 24 of switch 22 is connected to the input/output ports of disk devices 14. Disk identifier (ID) 48 and switch state (SS) 50 may control both the adapter 18 and switch 22 to establish appropriate connection between processor 12 and disk device 14. See column 3, line 48 to column 5, line 39; and Figure 1.

This reference relates to the use of a switch with a binding mapper and an address mapper. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an

exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

2. U.S. Patent No. 6,587,919 B2

The patent to Yanai et al. (6,587,919 B2), assigned to EMC Corp., provides for a System and Method for Disk Mapping and Data Retrieval. Disclosed is disk storage system 10 with means for receiving write commands and data such as channel adapter boards 12a-d, which receive disk read/write commands and data over communication channels 1-8. Disk adapter boards 20 read and write data to one or more disk drive units 22. Channel adapter boards 12a-d are connected to cache memory storage unit 16. When channel adapter boards 12a-d receive write commands, there is stored in memory an indication to disk adapters 20 that data record stored in cache must be written to the disk drives. See column 5, line 52 to column 6, line 18.

This reference relates to the use of a record locator data structure having variable-length data records for disk mapping and data retrieval. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

3. U.S. Patent No. 6,640,281 B2

The patent to Obara et al. (6,640,281 B2), assigned to Hitachi, Ltd., provides for Storage Subsystem with Management Site Changing Function. Disclosed is disk controller 8, host interfaces 2, channel paths 9, cache memory 4, disk interfaces 5, and plurality of disk drives 6. Cache memory temporarily stores data written in disk controller and data read from disk drive 8 and output to the host. Data exchange between host computer 1 and disk controller is performed via channel path 9, where data is called a frame. Frame received at host interface 2 is identified for target volume and type of operation from the volume number field 24 and command/data field 25, where the input/output command is sent

to head disk assemblies (HDA) of disk drive. For read command, HDA returns the read data to the disk interface. See column 4, line 30 to column 6, line 20; and Figures 1 and 2.

This reference relates to a technique for transferring management of desired disk drives and volumes to be managed by an overload disk controller under access concentration to an optional disk controller not in an overload state while an application on the host and an ordinary process of the disk controllers are maintained to continue. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

4. U.S. Patent No. 6,701,411 B2

The patent to Matsunami et al. (6,701,411 B2), assigned to Hitachi, Ltd., provides for a Switch and Storage System for Sending an Access Request from a Host to a Storage Subsystem. Disclosed is host adapter 101, diskarray interface (I/F) controller 1011 to control the diskarray switches 20, diskarray I/F 21, and host bus 1012 to perform communications and data transfer between cache memory 102 and diskarray I/F controller 1011. Lower adapter 103 executes control of disk I/F controller 1031 to control disk 104 and disk I/F. Diskarray switch 20 contains managing processor (MP) to perform functions such as management and control of diskarray switch and crossbar switch 201. During a read operation, switching controller (SC) 2022 reads the frame held in frame buffer (FB) 2021 and analyzes the frame header 401. Information such as ID for the frame transfer destination are included in the frame header. See column 4, line 55 to column 5, line 17; column 9, lines 27-45; and Figures 1 and 2.

This reference relates to the use of a switch connected between a first interface node and a plurality of second interface nodes to perform frame transfer between the first interface node and the second interface nodes based on node address information added to the frame. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of

the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

5. U.S. Patent Publication No. 2003/0191891 A1

The patent application publication to Tanaka et al. (2003/0191891 A1), assigned to Hitachi, Ltd., provides for a Disk Storage System having Disk Arrays Connected with Disk Adaptors through Switches. Disclosed is disk storage system comprising disk adapter DKA performing control required when data is transmitted and received between disk controller DKC and disk array DA, wherein DKA is connected to DA through channels D01-04. C1-4 are channels allowing communication between channel adapters CHA and CPU. Cache memory CM functions to temporarily store data inputted/outputted by disk storage system. Switch SW1 includes input/output ports P1-P5. During writing of data in DA, circuit configuration allows a frame in a block to be inputted from port P1 and outputted from ports P2-5. See paragraphs [0071], [0073], [0074], [0075], [0092], [0097], and [0098].

This reference discloses the use of switches between a disk adapter and a disk array to change over connection between ports to which the disk adapter is connected and ports to which disk drives constituting the disk array are connected in accordance with destination information in a frame for each of the inputted frames. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

6. U.S. Patent Publication No. 2005/0027919 A1

The patent application publication to Aruga (2005/0027919 A1) provides for a Disk Subsystem. Disclosed is N disk array controllers 1-1 to 1-N, each with M disk drive interface controllers 2-1 to 2-M, where each of the M controllers of fibre channel fabric switch 3-1 to 3-M are respectively connected to the disk drive interface controllers 2-1 to 2-

M for controlling disk drive units through fibre channel interface 5. Data transferred between host computer and disk array are temporarily stored in cache memory. Switch controller 17 sets switch 18 based on the ID number received from disk drive interface controllers 2-1 to 2-M. See paragraphs [0023], [0025], and [0029]; and Figure 1.

This reference relates to the use of a protocol controller disposed between switches in a fiber channel fabric switch circuit and disk drive units for converting a protocol to enable one-to-one connectivity established between controllers and disk drive units. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

7. Japanese Patent Publication No. JP 2003-85117

This reference provides a storage controlling device capable of efficiently using a full duplex type communication passage. The storage controlling device connects to an upward processing device through the full duplex type communication passage and stores and manages data received through the communication passage into a data storing means. The storage controlling device has a plurality of channel processors for inputting or outputting data from a data storing means in response to a command included in the data (frame) transmitted from the upward processing device, and assigns the channel processors for inputting or outputting the data related to the data (frame) in response to the command included in the data (frame).

As described in the present application at page 3, line 18 to page 4, line 2, the reference discloses that channel processors for inputting data to and outputting data from the disk device are controlled in accordance with a command from the host device and the quantity of data to be transferred so that full duplex operation is performed between the host device and the storage controlling device. However, dynamic control is required when data is transferred and its problem is complexity of the control method. Also, the reference does not deal with the full duplex data transfer in the back-end of a disk device. See page 5, lines 17-21.

This reference does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

8. Japanese Patent Publication No. JP 2000-222339

This reference discloses a way to connect plural disk drives with a disk drive interface circuit without sacrificing transmitting performance by using a fiber channel fabric topology for reducing the number of connection lines by using a fiber channel interface as a serial interface, and for realizing switch connection. A fiber channel switch control circuit 3 is provided between a disk drive 4 and a disk drive interface control circuit 2, and protocol control part 16 is provided between a switch 18 in the fabric switch circuit 3 and the disk drive.

As described in the present application at page 4, lines 3-8, the reference discloses a disk array system where a disk array controller and disk drives are connected via a switch. The reference, however, does not deal with application of the technique to the back-end of a disk drive equipped with a plurality of I/O ports and the full duplex data transfer in the back-end. See page 5, lines 22-25.

This reference does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

9. Japanese Patent Publication No. JP 2003-303055

This reference discloses a disk storage system having high throughput between a disk adapter of a disk controller and a disk array. The disk adapter of the disk controller is connected to the disk array through switches. Data on a channel between the switch and a RAID group is multiplexed in the switch to be transferred onto a channel between the switch

and the disk adapter, and data on the channel between the switch and the disk adapter is demultiplexed in the switch to be transferred onto the channel between the switch and the RAID group. A data transfer rate on the channel between the disk adapter and the switch is made higher than that on the channel.

This reference relates to the use of a switch with multiplexing and demultiplexing to achieve higher throughput between a disk adapter of a disk controller and a disk array. It does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

10. Qlogic Corp., "Full Duplex and Fibre Channel, Network Storage Group Host Products Technology Brief."

This reference discusses the meaning of full-duplex data transmission with fibre channel.

As discussed in the present application at page 3, lines 5-17, the reference discloses a plurality of FC-ALs in which disk drives are connected and a server are connected via a switch and parallel data transfers are carried out between the server and the plurality of FC-ALs. It does not take a disk drive having a plurality of I/O ports into consideration and it is difficult to apply the technique to a disk device comprising disk drives each having a plurality of I/O ports in the back-end. See page 5, lines 10-16.

This reference does not teach determining (1) a destination drive port to which a frame is to be forwarded or (2) a path which a frame passes to be transferred between a switch and one of the disk drives or (3) destination information within a frame to be transferred from the disk adapter to one of the disk drives, according to the type of a command included in an exchange that is transferred between the disk adapter and one of the disk drives, as recited in independent claims 1, 4, 6, 7, 8, and 9.

Appl. No. 10/770,723
Petition to Make Special

PATENT

(f) In view of this petition, the Examiner is respectfully requested to issue a first Office Action at an early date.

Respectfully submitted,



Chun-Pok Leung
Reg. No. 41,405

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400
Fax: 415-576-0300
Attachments
RL:rl
60475672 v1

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-085117

(43)Date of publication of application : 20.03.2003

(51)Int.Cl.

G06F 13/10

G06F 3/06

G06F 13/12

(21)Application number : 2001-273932

(71)Applicant : HITACHI LTD

(22)Date of filing : 10.09.2001

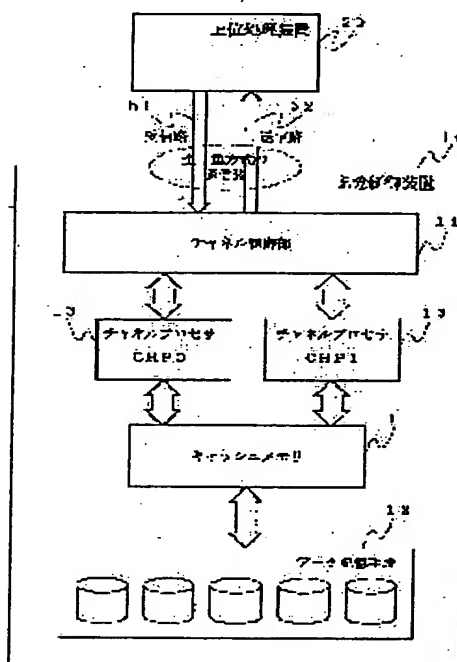
(72)Inventor :
MAEDA MASAMI
AZUMI YOSHIHIRO
SAKAKI TOSHINORI
TSUKADA MASARU

(54) STORAGE CONTROLLING DEVICE, AND ITS OPERATING METHOD

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a storage controlling device capable of efficiently using a full duplex type communication passage.

SOLUTION: This storage controlling device connects to an upward processing device through the full duplex type communication passage and stores and manages data received through the communication passage into a data storing means. The storage controlling device has a plurality of channel processors for inputting or outputting data from a data storing means in response to a command included in the data (frame) transmitted from the upward processing device, and assigns the channel processors for inputting or outputting the data related to the data (frame) in response to the command included in the data (frame).



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2003-85117

(P2003-85117A)

(43)公開日 平成15年 3月20日 (2003. 3. 20)

(51)Int.Cl.⁷

識別記号

F I

テマコード* (参考)

G 0 6 F 13/10

3 4 0

G 0 6 F 13/10

3 4 0 A 5 B 0 1 4

3/06

3 0 1

3/06

3 0 1 A 5 B 0 6 5

13/12

3 4 0

13/12

3 4 0 C

審査請求 未請求 請求項の数14 O L (全 14 頁)

(21)出願番号

特願2001-273932(P2001-273932)

(22)出願日

平成13年 9月10日 (2001. 9. 10)

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目 6 番地

(72)発明者 前田 昌美

神奈川県小田原市中里322番地 2 号 株式

会社日立製作所 R A I D システム事業部内

(72)発明者 安積 義弘

神奈川県小田原市中里322番地 2 号 株式

会社日立製作所 R A I D システム事業部内

(74)代理人 100071283

弁理士 一色 健輔 (外 5 名)

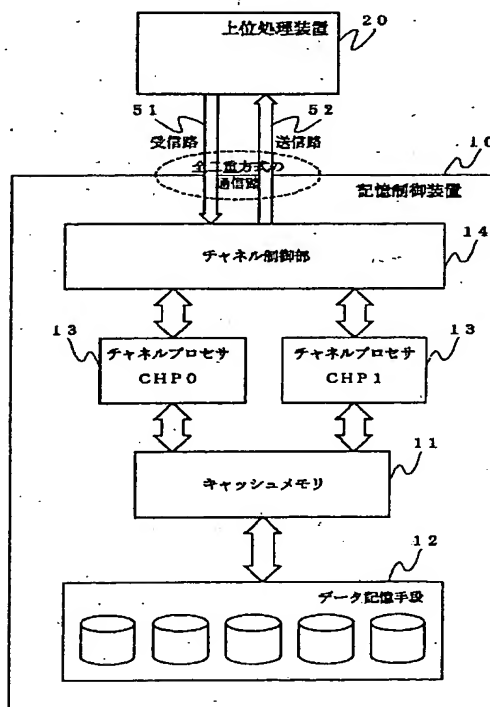
最終頁に続く

(54)【発明の名称】 記憶制御装置およびその運用方法

(57)【要約】

【課題】 全二重方式の通信路を効率よく利用することができる記憶制御装置を提供する。

【解決手段】 全二重方式の通信路により上位処理装置と接続し、通信路を通じて受信したデータをデータ記憶手段に記憶管理する記憶制御装置において、上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応してデータ記憶手段に対するデータ入出力処理を行う複数のチャネルプロセサを備え、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャネルプロセサを割り当てるようにする。



【特許請求の範囲】

【請求項 1】 全二重方式の通信路により上位処理装置と接続し、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理する記憶制御装置であって、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャンネルプロセッサを備え、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャンネルプロセッサを割り当てる手段を備えることを特徴とする記憶制御装置。

【請求項 2】 請求項 1 に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関するデータ入出力処理を実行する前記チャンネルプロセッサを割り当てる前記手段が、前記データ（フレーム）に含まれる前記コマンドが前記データ記憶手段にデータの書き込みコマンドであるかデータの読み出しコマンドであるかに応じて前記データ（フレーム）についての処理を行うチャンネルプロセッサを割り当てる手段であることを特徴とする記憶制御装置。

【請求項 3】 請求項 1 に記載の記憶制御装置であって、前記通信路の送信路と受信路を流れるデータ量に応じて前記データ（フレーム）についての処理を行うチャンネルプロセッサを割り当てる手段を備えることを特徴とする記憶制御装置。

【請求項 4】 請求項 1 に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータフレームに関するデータ入出力処理を実行する前記チャンネルプロセッサを割り当てる前記手段を実行するかどうかを、前記データ記憶手段に対するデータの書き込みコマンドとデータの読み出しコマンドにより単位時間内に処理されたデータ量に応じて制御する手段を備えることを特徴とする記憶制御装置。

【請求項 5】 請求項 1 に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関するデータ入出力処理を実行する前記チャンネルプロセッサを割り当てる前記手段を実行するかどうかを、前記各チャンネルプロセッサの処理待ちキューにキューイングされているデータ書き込みコマンドの数とデータ読み出しコマンドの数に応じて制御する手段を備えることを特徴とする記憶制御装置。

【請求項 6】 請求項 1 に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関するデータ入出力処理を実行する前記チャンネルプロセッサを割り当てる前記手段を実行するかどうかを、前記各チャンネルプロセッサが前記データ入出力処理において単位時間内に処理したデータ量に応じて制御する手段を備えることを特徴とする

記憶制御装置。

【請求項 7】 請求項 1 に記載の記憶制御装置であって、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関するデータ入出力処理を実行する前記チャンネルプロセッサを割り当てる前記手段を実行するかどうかを、当該記憶制御装置における、前記データ記憶手段に対するデータ書き込みコマンドの処理についてのスループットと、前記データ記憶手段に対するデータ読み出しコマンドの処理についてのスループットに応じて制御する手段を備えることを特徴とする記憶制御装置。

【請求項 8】 請求項 4 または 6 に記載の記憶制御装置であって、前記単位時間を当該記憶制御装置に接続された外部装置から指定させる手段を備えることを特徴とする請求項 4 または 6 に記載の記憶制御装置。

【請求項 9】 全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャンネルプロセッサを備えて構成される記憶制御装置の運用方法であって、前記データ（フレーム）に含まれる前記コマンドが前記データ記憶手段にデータの書き込みコマンドであるかデータの読み出しコマンドであるかに応じて前記データ（フレーム）についての処理を行うチャンネルプロセッサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項 10】 全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャンネルプロセッサを備えて構成される記憶制御装置の運用方法であって、前記通信路の送信路と受信路を流れるデータ量に応じて前記データ（フレーム）についての処理を行うチャンネルプロセッサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項 11】 全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャンネルプロセッサを備えて構成される記憶制御装置の運用方法であって、前記データ（フレーム）に含まれるコマンドの種類と、前記データ記憶手段に対するデータの書き込みコマンドとデータの読み出しコマンドにより単位時間内に処理されたデータ量とに応じてそのデータ（フレーム）に関す

る前記データ入出力処理を実行するチャンネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項 12】 全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャンネルプロセサを備えて構成される記憶制御装置の運用方法であって、

前記データ（フレーム）に含まれるコマンドの種類と、前記各チャンネルプロセサの処理待ちキューにキューイングされているデータ書き込みコマンドの数とデータ読み出しコマンドの数とに応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャンネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項 13】 全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャンネルプロセサを備えて構成される記憶制御装置の運用方法であって、

前記データ（フレーム）に含まれるコマンドの種類と、前記各チャンネルプロセサが前記データ入出力処理において単位時間内に処理したデータ量とに応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャンネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【請求項 14】 全二重方式の通信路により上位処理装置に接続され、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理し、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャンネルプロセサを備えて構成される記憶制御装置の運用方法であって、

前記データ（フレーム）に含まれるコマンドの種類と、当該記憶制御装置における、前記データ記憶手段に対するデータ書き込みコマンドの処理についてのスループットと、前記データ記憶手段に対するデータ読み出しコマンドの処理についてのスループットとに応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャンネルプロセサを割り当てるようにしたことを特徴とする記憶制御装置の運用方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 この発明は、全二重方式の通信路により上位処理装置と接続し、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理する記憶制

御装置に関し、とくに、前記通信路を効率よく利用するための技術に関する。

【0002】

【従来の技術】 メインフレームなどの上位処理装置と、ディスクアレイ装置などの記憶制御装置との間を結ぶ通信プロトコルとして、ファイバチャネルプロトコル（以下、「ファイバチャネル」と称する）が知られている（詳細はANSI (American National Standard for Information Technology) のFC-PH (Fibre Channel Physical and Signaling Protocol) を参照）。

【0003】 ファイバチャネルは、2ポート間の物理接続が基本であり、一対のファイバチャネルポートは、物理的には送受信2本（全二重方式）の通信路で相互に接続されている。この通信路を通じて行われるファイバチャネルにおける記憶制御装置と上位処理装置との間の通信は、フレームと呼ばれるデータ単位を基本として行われる。また、複数のフレームの束はシーケンスと呼び、さらに、シーケンスの束はエクステンジと呼ぶ。例えば、上位処理装置から記憶制御装置に対してデータ読み出し命令（Read命令）に対応する一連の処理はエクステンジを単位として行われる。また、ファイバチャネルにおける上位処理装置と記憶制御装置との間の通信では、インターロックを取らずにコマンドやフレームを送受信することが可能である。

【0004】 図1に、ファイバチャネルの通信路により接続された記憶制御装置10と、これを利用する上位処理装置20とを備えて構成されるデータ処理システムの一例を示す。記憶制御装置10は、例えば、ディスクアレイ装置などであり、キャッシュメモリ11、ディスクユニットなどのデータ記憶手段12、CPUやメモリなどを主体として構成される。そして、記憶制御装置10は、上位処理装置20との間のデータ通信や処理対象となるデータのチャンネルプロセサ13への割り当ておよび各種データやテーブルの管理や、チャンネルプロセサ13への各種命令のキューの管理などを行うチャンネル制御部14、チャンネル制御部14から受信したフレームを切り分けて各フレームに対応する処理やキャッシュメモリ11を通じたデータ記憶手段12へのデータ転送制御を実行するチャンネルプロセサ13などを備えている。一方、上位処理装置20は、例えば、メインフレームやオフコン、パーソナルコンピュータなどである。

【0005】 記憶制御装置10と上位処理装置20との間の通信において、記憶制御装置10は、受信したフレームの順に各フレームを処理するチャンネルプロセサ13を割り当てる。また、この際にチャンネルプロセサ13が使用中の場合には、使用されていないチャンネルプロセサ13がそのフレームの処理用に割り当てられることになる。

【0006】

【発明が解決しようとする課題】 ところで、上位処理装

置20から記憶制御装置10に送られてくるフレームに含まれるコマンドは、主としてデータ記憶手段12へのデータの書き込みを指示するWRITEコマンドと読み出しを指示するREADコマンドに大別され、これらコマンドに対応する処理に際しては、WRITEコマンドもしくはREADコマンドの一方の種類のコマンドのみを含んだフレームが連続して上位処理装置20から送られてきた場合や、書き込みや読み出しの対象となるデータ量が大きいなどの理由により、複数のチャンネルプロセッサ13の双方が同時にWRITEコマンドもしくはREADコマンドのどちらか一方の処理のみを行っている期間が生じることがあり、この期間中は全二重通信が有効に機能しないことになる。

【0007】すなわち、例えば、チャンネルプロセッサ13が2つのみで構成されている場合には、前述した全二重方式の通信路のうち上位処理装置20から記憶制御装置10方向の通信路51（以下、「受信路」と称する）のみの通信負荷が高くなっているにもかかわらず、記憶制御装置10から上位処理装置20方向の通信路52（以下、「送信路」と称する）は殆ど利用されていないため、この期間中は全二重方式の通信路50が有効に機能しないことになる。

【0008】本発明はこのような事情に鑑みてなされたもので、全二重方式の通信路を効率よく利用することができる記憶制御装置を提供することを目的とする。

【0009】

【課題を解決するための手段】この目的を達成するための、本発明の主たる発明は、全二重方式の通信路により上位処理装置と接続し、前記通信路を通じて受信したデータをデータ記憶手段に記憶管理する記憶制御装置であって、前記通信路を通じて上位処理装置から送られてくるデータ（フレーム）に含まれるコマンドに対応して前記データ記憶手段に対するデータ入出力処理を行う複数のチャンネルプロセッサを備え、前記データ（フレーム）に含まれるコマンドの種類に応じてそのデータ（フレーム）に関する前記データ入出力処理を実行するチャンネルプロセッサを割り当てる手段を備えることとする。

【0010】具体的には、例えば、ファイバチャネルにより接続された上位処理装置と記憶制御装置との通信の場合であれば、前述のインターロックが不要であるというファイバチャネルの特徴を利用して、フレームに含まれているコマンドが前記データ記憶手段に対する書き込み命令であるのか、読み出し命令であるのかに応じてそのフレームが所属するエクステンジを処理するチャンネルプロセッサをエクステンジごとに割り当てる。これにより、全てのチャンネルプロセッサが同時に書き込みもしくは読み出しの一方の処理のみを実行する期間が減り、これにより全二重方式の通信路の送信路と受信路の負荷バランスの均一化が図られ、通信路を効率よく利用することが可能になる。

【0011】

【発明の実施の形態】以下、本発明の一実施例によるデータ処理システムについて説明する。データ処理システムの構成は図1と同様であるので詳細な説明は省略し、以下では本発明の特徴的な部分を中心に説明する。また、説明の便宜のため、この実施例で説明する記憶制御装置10は、2つのチャンネルプロセッサ（以下、「CHP」と称する）13のみを備えるものとし、これらチャンネルプロセッサをCHP0、CHP1と称することとする。また、以下の説明では「WRITE」を「WR」と、「READ」を「RD」と、それぞれ略記する。

【0012】記憶制御装置10のチャンネル制御部14のメモリ上には、上位処理装置20との間のデータ通信や、データ記憶手段12に対するデータの読み書き処理などの、上位処理装置20から送られてくるコマンドに応じて行われる処理に利用される各種テーブルが記憶管理されている。図2はこのうちのエクステンジ管理テーブル200である。このテーブル200には、上位処理装置20との間の通信において生成されたエクステンジが管理されている。

【0013】このテーブルのテーブル有効フラグ202には、該当のエクステンジが現在処理中であるかどうかを示すビットがセットされ、処理中のエクステンジについては「1」が、処理中でないエクステンジについては「0」がセットされる。デバイス番号203には、そのエクステンジの処理対象となるデータ記憶手段の記憶媒体（例えば、ディスクユニット）の識別番号がセットされる。CHP番号205には当該エクステンジの処理を行うCHPの番号がセットされる。また、通信領域ポインタ204には受信路51を通じて受信したフレームが格納されているメモリ上の物理アドレスが、OX-ID (Originator Exchange_ID) 207には、上位処理装置20側で割り当てられたエクステンジ番号がセットされ、S-ID (Source_ID) 208には送信元のファイバチャネルのポートアドレスが、D-ID (Destination_ID) 209には該当エクステンジの送信先となるファイバチャネルのポートアドレスが、それぞれセットされる。

【0014】なお、以上の項目のうちテーブル有効フラグ以外の項目には、エクステンジが処理中である場合、すなわち、テーブル有効フラグが「1」の場合に値がセットされる。また、OX-ID 207、S-ID 208、D-ID 209の値は、各エクステンジに一意に対応づけられ、例えば、上位処理装置20から連続してフレームが送られてきた場合に各フレームがどのエクステンジに所属しているか、フレームが先頭フレームであるか（エクステンジ管理テーブルにフレームが登録されていない場合）といったことは、各フレームにセットされているOX-ID 207、S-ID 208、D-ID 209に一致するエクステンジをエクステンジ管理テーブル200から調べることで把握できる。

【0015】一方、図3は記憶制御装置10のメモリ上に記憶管理されているCHP割当処理管理テーブル300である。このテーブルのCHP割当処理実行フラグ301は、CHPの割当処理の実行制御に用いられる。このフラグの用途については後述する。

【0016】受信WRコマンドデータ量カウンタ302および受信RDコマンドデータ量カウンタ303には、チャネル制御部14が上位処理装置20からコマンドを受信した場合に、WRコマンドもしくは、RDコマンドにより処理されたデータ量が加算される。また、送信RDコマンドデータ量カウンタ304には、上位処理装置20から記憶制御装置10に送信されたRDコマンドにตอบสนองして記憶制御装置10から上位処理装置20に送信されるデータ量がセットされ、CHP13から処理完了通知があり、あるエクスチェンジがエクスチェンジ管理テーブル200において無効化された際に値が加算される。

【0017】CHP0データ量カウンタ305、およびCHP1データ量カウンタ306には、各CHP0、1が処理したデータのデータ量が加算される。RD/WRデータ比率閾値等の各種閾値307~311は、後述するCHPの割り当て処理の実行制御のために利用されるものである。モニタリング実行間隔タイマ312、モニタリング実行間隔I/O数313は、記憶制御装置10が実行する各種ポーリング処理に際して参照されるパラメータである。また、モニタリング開始時刻314とモニタリング終了時刻315は、記憶制御装置10により処理されたデータのスループットを算出する場合に用いられる。I/O数カウンタ315は、I/O数でモニタリングするのに必要な情報であり、これには上位処理装置20から送られてきたI/O要求数が加算される。ワークエリア316は、各種計算やデータの一時的な保存などに利用される。

【0018】図4は、チャネル制御部14のメモリ上に記憶管理されているRD/WRコマンドキュー管理テーブル401、410である。このテーブル401、410には、記憶制御装置20の処理対象となるRDコマンドもしくはWRコマンドのキューイング状態が、FIFO (First In First Out) 方式で管理され、キューイング数を示すカウンタ402、先頭キューの格納アドレスを示すINポインタ403、末尾キューの格納アドレスを示すOUTポインタ404、キュー全体のデータサイズを管理するデータ転送量405、コマンドごとに対応するキュー管理情報406などが管理される。キュー管理情報406には、記憶制御装置20が割り当てたエクスチェンジ番号407とデータ転送量408、このエクスチェンジを実行するCHPの番号409などが記述されている。

【0019】つぎに、図5に示すフローチャートとともに、記憶制御装置10と上位処理装置20との間の全二重方式の通信において記憶制御装置10が行う、CHP

の割当処理について詳述する。

【0020】全二重方式の通信路のうち受信路51を通じて上位処理装置20からフレームを受信した場合、記憶制御装置10は、まず、そのフレームが新たなエクスチェンジの起動先頭フレームであるかどうかを調べる(501、502)。ここでフレームが起動先頭フレームであった場合には、未使用のエクスチェンジ番号を利用してそのフレームに対応するフィールドをエクスチェンジ管理テーブルに新たに登録する(503)。なお、フレームが起動先頭フレームでない場合の処理については後述する。

【0021】つぎに記憶制御装置10は、前記フレームのフレーム制御フィールド(F_ControlField)を参照し、当該フレームの後続フレームの存在有無を調べる(507)。その結果、当該フレームが後続フレームを有する場合には、さらにこのフレームがRD/WRいずれかのコマンドを含むフレームであるかどうかを調べる(508)。なお、この調査は、例えば、記憶制御装置10内にあらかじめ登録しておいたコマンド一覧と、フレームのコマンド記述欄の内容を比較することで行う。

【0022】この調査の結果、このフレームがコマンドも含んでいない場合(もしくは、コマンドを含んでいるかどうかを判断できない場合)には、記憶制御装置10は、当該フレームを通常のCHP割当方式、すなわち、受信したフレームの順に各フレームを処理するチャネルプロセッサ13を割り当て、また、あるフレームの割り当てに際してチャネルプロセッサ13が使用中である場合には、使用されていないチャネルプロセッサ13にそのフレームの処理させるというCHP割当方式により当該フレームの処理を行う(510)。

【0023】一方、前記フレームがRD/WRいずれかのコマンドを含むフレームである場合には、つぎのように処理が行われる。まず、フレームにWRコマンドが含まれていた場合、記憶制御装置10は、WRデータキュー管理テーブル401に当該フレームのエクスチェンジ番号とそのWRコマンドにより処理されるデータ量を登録する(511)。一方、当該フレームにRDコマンドが含まれていた場合には、記憶制御装置はRDデータキュー管理テーブルに当該フレームのエクスチェンジ番号と、そのREADコマンドにより処理されるデータ量を登録する(512)。なお、これら各キュー管理テーブルへの登録が行った場合には、各テーブルのカウンタ402およびOUTポインタ404に1を加算する。

【0024】つぎに記憶制御装置10は、条件管理テーブルのCHP割当処理実行フラグ301の状態を調査する。ここでCHP割当処理実行フラグ301に「1」がセットされていた場合には、CHPの割当処理を行うかどうかの判断する処理に進み(513、514)、当該フレームにWRコマンドが含まれていた場合は、エクスチェンジ管理テーブル200の当該フレームが所属するエ

クスチェンジのCHP番号205の欄に、そのエクスチェンジにCHP0を割り当てたことを示す「0」をセットし(515)、また、CHP割当処理管理テーブル300のCHP0データ量カウンタ305に当該フレームのデータ量を加算し(516)、当該フレームの処理を行うため当該フレームについての処理に必要なデータをCHP1に転送する(519)。なお、フレームにRDコマンドが含まれていた場合も、以上のWRコマンドの場合と同じようにして処理が行われる(512, 514, 517, 518)。

【0025】一方、(513, 514)の処理において、CHP割当処理実行フラグ301に「0」がセットされていた場合には、記憶制御装置10はコマンドの種類(WRコマンドであるかRDコマンドであるか)に応じたCHPの割当処理を行わず、その代わりにCHP割当処理管理テーブル300のCHP0データ量カウンタ305もしくはCHP1データ量カウンタ306の比率に基づいてCHPの割り当てを行う(521)。すなわち、この場合、記憶制御装置10は、CHP0データ量カウンタ305およびCHP1データ量カウンタ306の値を比較して、値の小さい方、すなわち、その時点で処理負荷の小さいCHPを当該フレームの処理用に割り当て、当該フレームの処理に必要なデータをそのCHPに転送する。

【0026】つぎに、(502)の処理において、フレームが起動先頭コマンドでなかった場合には、まず、フレーム情報(OX-ID, S-ID, LPN番号、デバイス番号)をキーとした場合に該当するエクスチェンジをエクスチェンジ管理テーブル200から検索する。そして、検索したエクスチェンジのCHP番号205に既に値がセットされていれば、当該フレームの処理に必要なデータをその値に対応するCHPに送信する(516)。

【0027】他方、CHP番号205に値がセットされていない場合、すなわち、そのフレームが所属するエクスチェンジに既にCHPが割り当てられていない場合には、(507)からの処理に進む。以上のようにして記憶制御装置10は受信したフレームをつぎつぎに処理していくことになる。

【0028】つぎに、以上の処理をより具体的に説明すべく、上位処理装置20から記憶制御装置10に対し、DX/LOC/WRCKDからなる1CCW(Channel Command Word)チェーン(例えば、「IBM 3990/9390 Storage Control Reference」を参照)のフレームが送られてきた場合の処理について、再度、図5のフローチャートに従って説明する。

【0029】記憶制御装置10は、DXコマンドを含んだフレームを受信すると、まず、このフレームが起動先頭コマンドフレームかどうかを判定する(508)。ここでDXコマンドはCCWチェーンの起動先頭コマンドであ

り、新たなエクスチェンジの起動先頭フレームであるので、記憶制御装置10はこのフレームに対応するエクスチェンジを、エクスチェンジ管理テーブル200に新規に登録する。

【0030】つぎに記憶制御装置10は、前記フレームのフレーム制御フィールド(F_ControlField)を参照し、当該フレームの後続フレームの存在有無を調べる(507)。その結果、後続フレームを有する場合には、さらにこのフレームがRD/WRいずれかのコマンドを含むフレームであるかどうかを調べる(508)。ここでDXコマンドは、READ/WRITEいずれのコマンドでも無いため、記憶制御装置10は、当該フレームを通常のCHP割当方式により割り当てたCHPにより処理することになる(510)。

【0031】つぎに、記憶制御装置10は、DXコマンドに引き続きCCWチェーンを構成するLOCコマンドを含むフレームを受信すると、このフレームが起動先頭コマンドであるかどうかを調べる(502)。ここでLOCコマンドが記載されたフレームは起動先頭コマンドで無いため、記憶制御装置10はこのフレームのフレーム情報(OX-ID, S-ID, LPN番号、デバイス番号)を用いてこのフレームに対応するエクスチェンジをエクスチェンジ管理テーブル200より検索する。そして、この場合には、前述のDXコマンドが記載されたフレームにより前記フレーム情報に対応するエクスチェンジが既にエクスチェンジ管理テーブル200に登録されているため、検索の結果、このエクスチェンジが検索されることになる。

【0032】つぎに、記憶制御装置10は、検索されたエクスチェンジについて、当該エクスチェンジの処理用に既にCHPが割り当てられているかどうかを当該エクスチェンジのCHP番号205を参照して調べる(506)。ここでこのエクスチェンジには、まだCHPの割り当てがされていないので、(507)の処理において当該フレームの後続チェーンが存在するかどうかを調べる。ここでこのフレームには後続チェーン(WRCKDコマンドのフレーム)が存在するため、(508)の処理が実行される。そして、LOCコマンドは、そのオペレーションコードからWRITEコマンドであることを判定できるコマンドであるため、(510)の処理へと移行して、当該LOCコマンドをWRデータキュー管理テーブル410に登録し、また、エクスチェンジ管理テーブル200の当該フレームに対応するエクスチェンジのCHP番号205に「0」をセットする(511, 513)。また、CHP0データ量カウンタ305に当該LOCコマンドの処理されるデータ量、例えば、当該コマンドによりディスクユニットに書き込まれるデータのデータ量を加算する。

【0033】つぎに、上位処理装置20からWRCKDコマンドが記載されたフレームが送られてきた場合には、記

憶制御装置10はこのフレームは起動先頭フレームでないため(502)、前述したLOCコマンドの場合と同様に当該フレームが所属するエクスチェンジ番号を検索する(505)。そして、この場合には、CHP番号205に既に値がセットされており、当該フレームが所属するエクスチェンジに既にCHPが割り当てられているため(506)、記憶制御装置10は当該コマンドの処理に必要なデータをそのCHP番号に対応するCHPに送信する。

【0034】以上に説明したように、記憶制御装置10はフレームに含まれるコマンドがRDコマンドであるかWRコマンドであるかに応じてそのフレームが所属するエクスチェンジの処理を行うCHPを割り当てる。従って、CHP0、1の双方が同時にWRもしくはRDの一方の処理を行う期間が減って、全二重方式の通信路の受信路51と送信路52の負荷のバランスが不均一になる期間を減らすことができる。

【0035】ところで、以上のようなCHPの割当方式を適用した場合でも、例えば、上位処理装置20からWRコマンドもしくはRDコマンドのいずれか一方のみを含むフレームが連続して送信されて長期間一方のCHPがその処理に占有されていたり、一のコマンドの処理対象となるデータのデータ量が大きい場合に前記の割当処理を実行してしまうと、かえって一方の通信路に負荷が片寄る結果となり、通信路の負荷のバランスが崩れてしまう可能性がある。そこで負荷分散をより徹底して行うようにするため、本発明の記憶制御装置10はさらに以下に示すような各種の機能を備えている。

【0036】このうち第1の機能は、CHPの割当処理を行うかどうかを、上位処理装置20から送られてくるRDコマンドとWRコマンドのそれぞれにより処理されるデータ量の比率に応じて制御するようにしたものである。具体的には、図6に示すように、ポーリング処理などにより適宜なタイミングである一定期間におけるCHP割当処理管理テーブル300の受信WRコマンドデータ量カウンタ302と、受信RDコマンドデータ量カウンタ303の増分からRDコマンドとWRコマンドのそれぞれにより処理されたデータ量の比率を算出(602)し、この比率がCHP割当処理管理テーブル300のRD/WRデータ量比率閾値を超えた場合にはCHP割当処理管理テーブル300のCHP割当処理実行フラグ301に「0」をセットし(605)、RD/WRデータ量比率閾値以下の場合にはCHP割当処理実行フラグ301に「1」をセットする(604)。なお、図6の例では、ポーリング機能などにより以上の処理を実行するたびに、受信WRデータカウンタと、受信RDデータカウンタの内容を初期化している(607)。

【0037】第2の機能は、各CHPのキューの状態に応じてCHP割当処理実行フラグを制御するようにしたものである。具体的には、ポーリング処理などにより一

定期間ごともしくは処理データが一定数に達する度となどの適宜なタイミングで、RD/WRデータキュー管理テーブルから各CHP0、1にキューイングされているコマンドの数およびこれらコマンドの処理対象となるデータの全データ量を算出する。キューイングされているコマンド数はRD/WRキュー管理テーブルのカウンタの値により把握される。また、全データ量は、データ転送量により把握される。RDコマンドおよびWRコマンドそれぞれのデータ転送量の比率を、それぞれCHP割当処理管理テーブル300のCHPキューデータ数比率閾値309、CHPキューデータ量比率閾値310と比較し、閾値を超えているかどうかに応じてCHP割当処理実行フラグ301を制御する。図7は以上の処理の一例を示すフローチャートである。

【0038】第3の機能は、CHP割当処理管理テーブル300のCHP0、1データ量カウンタ305、306の比率に応じてCHP割当処理実行フラグ301を制御するようにしたものである。具体的には、一定時間毎にCHP0、1データ量カウンタ305、306の比率と、CHP割当処理管理テーブル300のCHPデータ比率閾値307とを比較し、閾値を超えているかどうかに応じてCHP割当処理実行フラグ301を制御する。図8にこの場合の処理の一例を示す。

【0039】第4の機能は、WRコマンドとRDコマンドのそれぞれについて、記憶制御装置10が単位時間当たり処理したデータ量(スループット)の比に応じてCHP割当処理実行フラグ301を制御するようにしたものである。ここでWRコマンドとRDコマンドそれぞれについてのスループットは、CHP割当処理管理テーブル300における受信WRコマンドデータ量カウンタ302および送信RDコマンドデータ量カウンタ304の変化率により算出する(901)。すなわち、図9に示すように、ある時刻におけるこれらカウンタ値とこれから単位時間経過後のこれらカウンタ値の差から単位時間当たりの処理データ量であるスループットを算出し(901~903、907、908)、このようにして求めたスループットの比がCHP割当処理管理テーブル300のRD/WRスループット比率閾値311を超えているかどうかに応じてCHP割当処理実行フラグ304を制御する(904)。

【0040】以上に説明した第1~第4の機能によれば、上位処理装置20からWRコマンドもしくはRDコマンドのいずれか一方のみが記載されたフレームが連続して送られてきたり、一のコマンドの処理対象となるデータのデータ量が大きい場合における一方の通信路に負荷の片寄りを防ぐことが可能となり、フレームに含まれるコマンドの種類(RDコマンドもしくはWRコマンド)に応じてそのフレームを処理するCHPを割り当てることで、全二重通信における受信路と送信路の負荷分散を図る前

述した仕組みをより一層効果的に機能させることが可能となる。なお、以上に説明した第1～第4の機能は、全てを一度に適用しなければならない訳ではなく、このうちの1の機能のみを適用したり、いくつかの機能を選択して適用するようにしてもよい。

【0041】また、以上に説明した第1～第4の機能において、ポーリング処理の間隔を指定する数値、例えば、時間や処理データ数などの数値を、上位処理装置20や、記憶制御装置10に接続された運用管理端末などの外部装置から指定できるようにしてもよい。

【0042】ところで、以上の実施例は、通信プロトコルがファイバチャネルプロトコルである場合について説明したが、上位処理装置20と2以上のチャネルプロセッサを備えた記憶制御装置10が全二重方式で結ばれる構成を備えるデータ処理システムであれば、通信プロトコルの種類に限定されることなく適用することができる。

【0043】以上の実施例は、記憶制御装置10が一つのチャネル制御部14に対し2つのCHP13を備えている場合について説明したが、一つのチャネル制御部14に対して3つ以上のCHP13を備えている場合にも適用できることはもちろんである。

【0044】

【発明の効果】以上に説明したように、本発明の記憶制御装置によれば、全二重方式の通信路を効率よく利用することができる。

【図面の簡単な説明】

【図1】本発明の一実施例によるデータ処理システムの概略構成を示す図である。

【図2】本発明の一実施例によるエクスチェンジ管理テーブルを示す図である。

【図3】本発明の一実施例によるCHP割当処理管理テ

ーブルを示す図である。

【図4】本発明の一実施例によるRD/WRコマンドキュー管理テーブルを示す図である。

【図5】本発明の一実施例によるCHP割当処理を説明するフローチャートである。

【図6】本発明の一実施例による、CHP割当処理を行うかどうかを、RDコマンドとWRコマンドのそれぞれにより処理されるデータ量の比率に応じて制御する処理を説明するフローチャートである。

【図7】本発明の一実施例による、CHP割当処理を行うかどうかを、各CHPのキューの状態に応じてCHP割当処理実行フラグを制御する処理を説明するフローチャートである。

【図8】本発明の一実施例による、CHP割当処理を行うかどうかを、CHP0、1データ量カウンタの比率に応じて制御する処理を説明するフローチャートである。

【図9】本発明の一実施例による、CHP割当処理を行うかどうかを、WRコマンドとRDコマンドのそれぞれについて、記憶制御装置が単位時間当たり処理したデータ量（スループット）の比に応じて制御する処理を説明するフローチャートである。

【符号の説明】

- 10 記憶制御装置
- 11 キャッシュメモリ
- 12 データ記憶手段
- 13 チャネルプロセッサ
- 14 チャネル制御部
- 20 上位処理装置
- 51 受信路
- 52 送信路

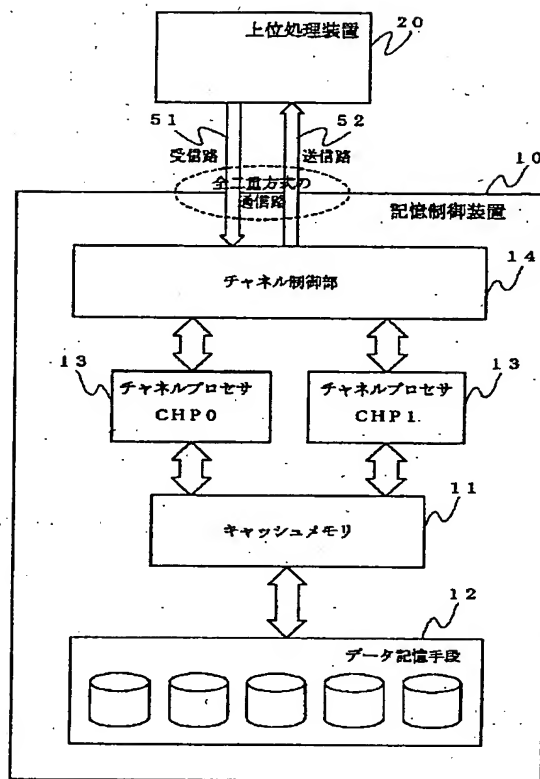
【図2】

201	202	203	204	205	206	207	208	209	210
EXID	トランスミッタ番号	デバイス番号	通信領域	CHP番号	処理データ	OX-ID	S-ID	D-ID	データ量
1	1	05	0x0000	FF	0001	0004	0100	0200	2045
2	1	09	0x0040	01	0001	0005	0300	0400	4096
3	1	01	0x0080	00	0003	0006	0500	0600	8192
4	0								
:									
:									
K	0								

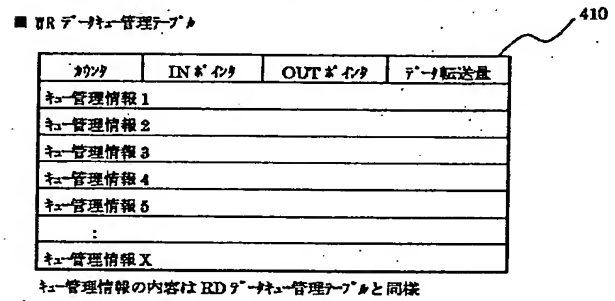
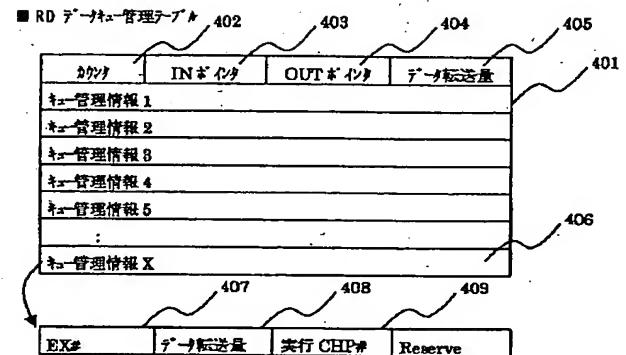
【図3】

CHP割当処理実行フラグ	301
受信WRコマンドデータ量カウンタ	302
受信RDコマンドデータ量カウンタ	303
送信RDコマンドデータ量カウンタ	304
CHP0データ量カウンタ	305
CHP1データ量カウンタ	306
CHPデータ比率閾値	307
RD/WRデータ比率閾値	308
CHPキューデータ数比率閾値	309
CHPキューデータ量比率閾値	310
RD/WRスループット比率閾値	311
モニタリング実行間隔時間	312
モニタリング実行間隔I/O数	313
モニタリング開始時刻	314
モニタリング終了時刻	315
I/O数カウンタ	316
ワークエリア	317

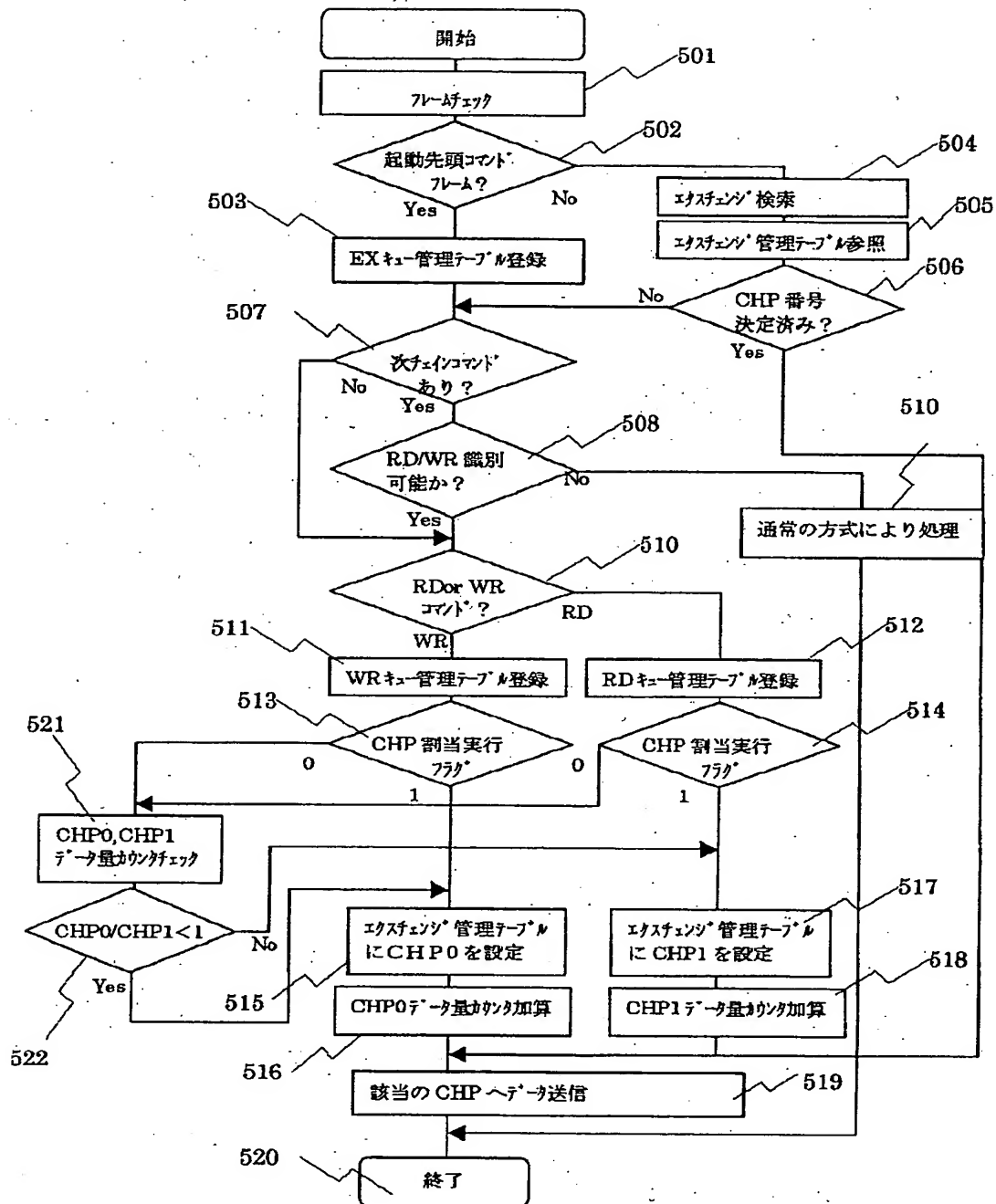
【図1】



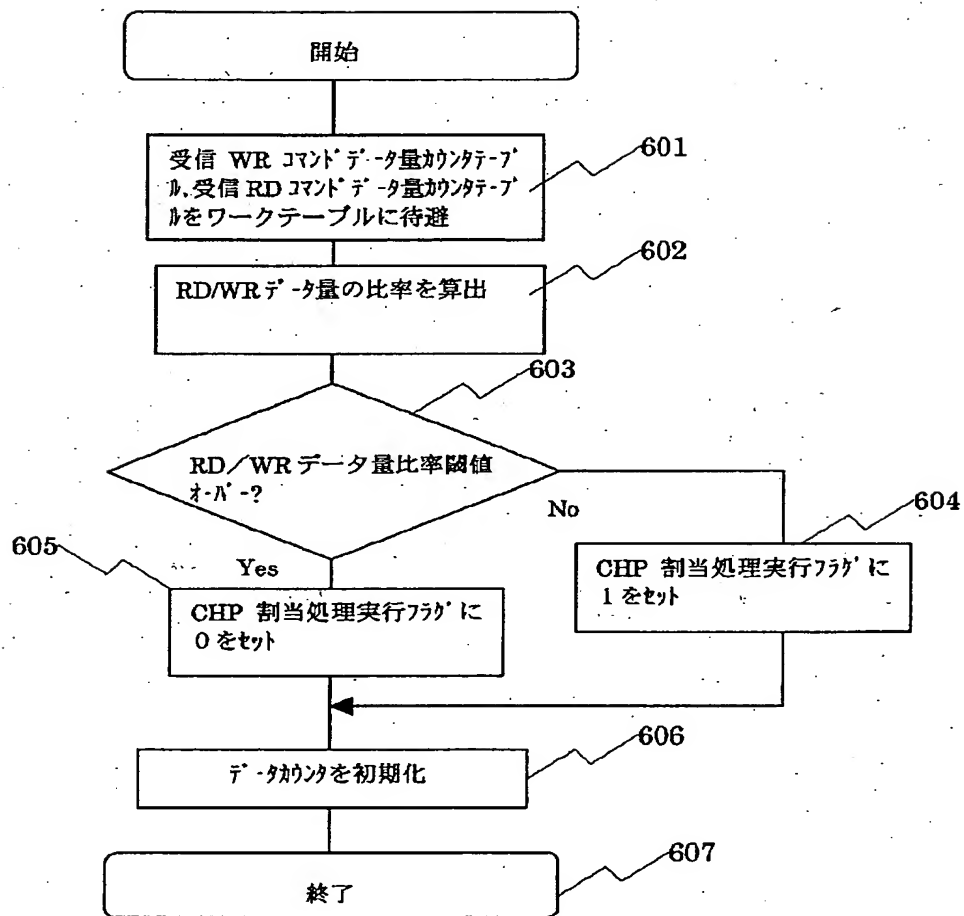
【図4】



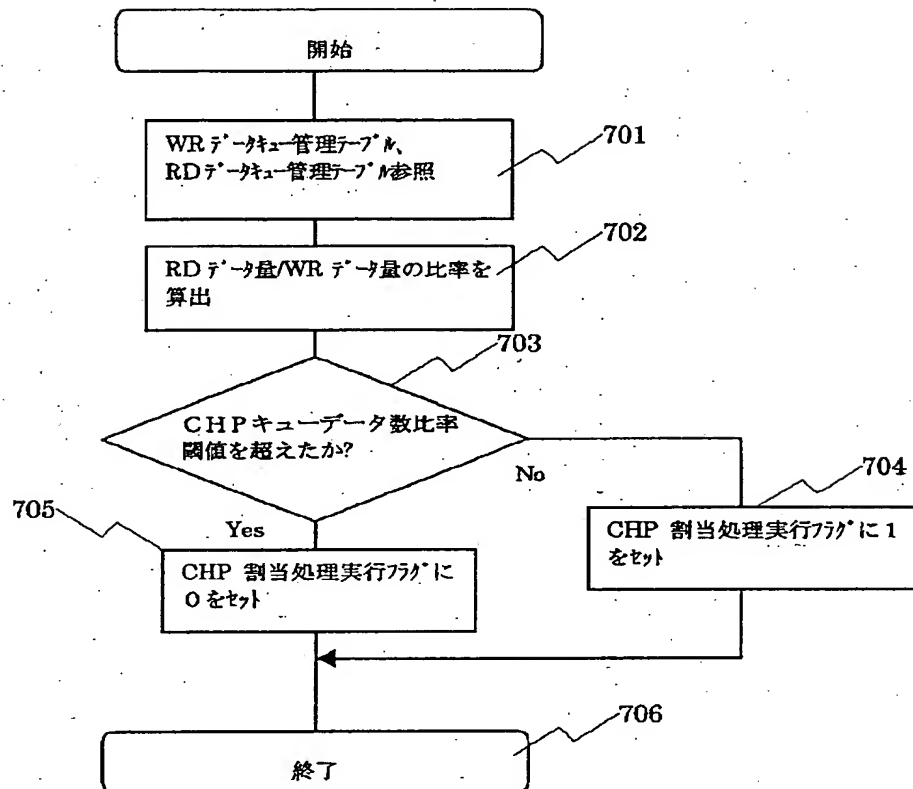
【図5】



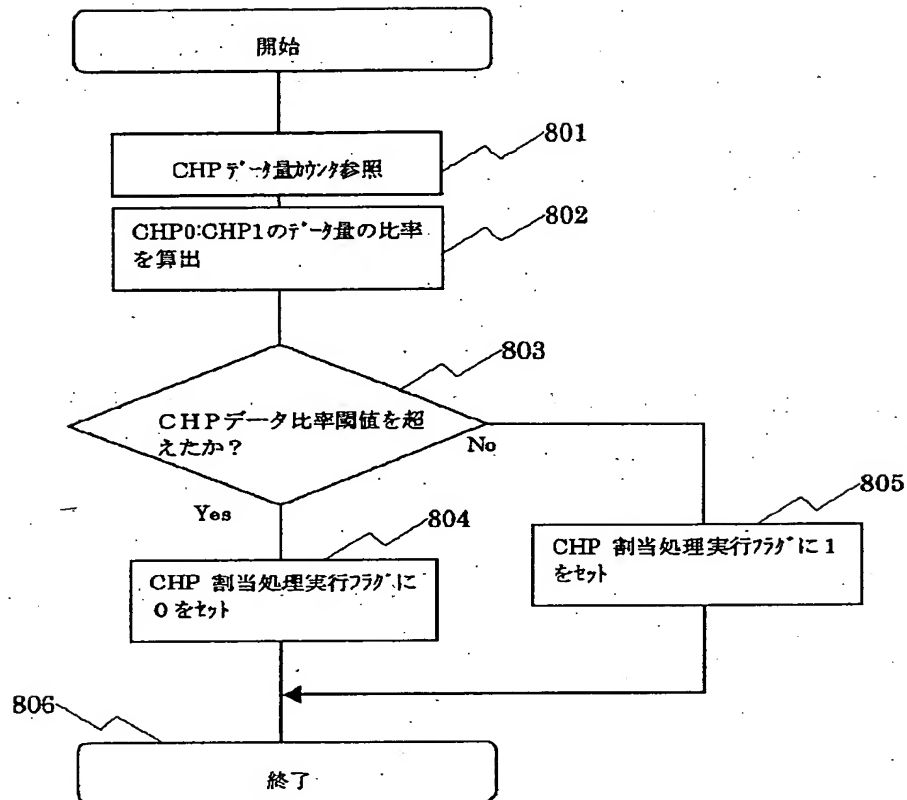
【図6】



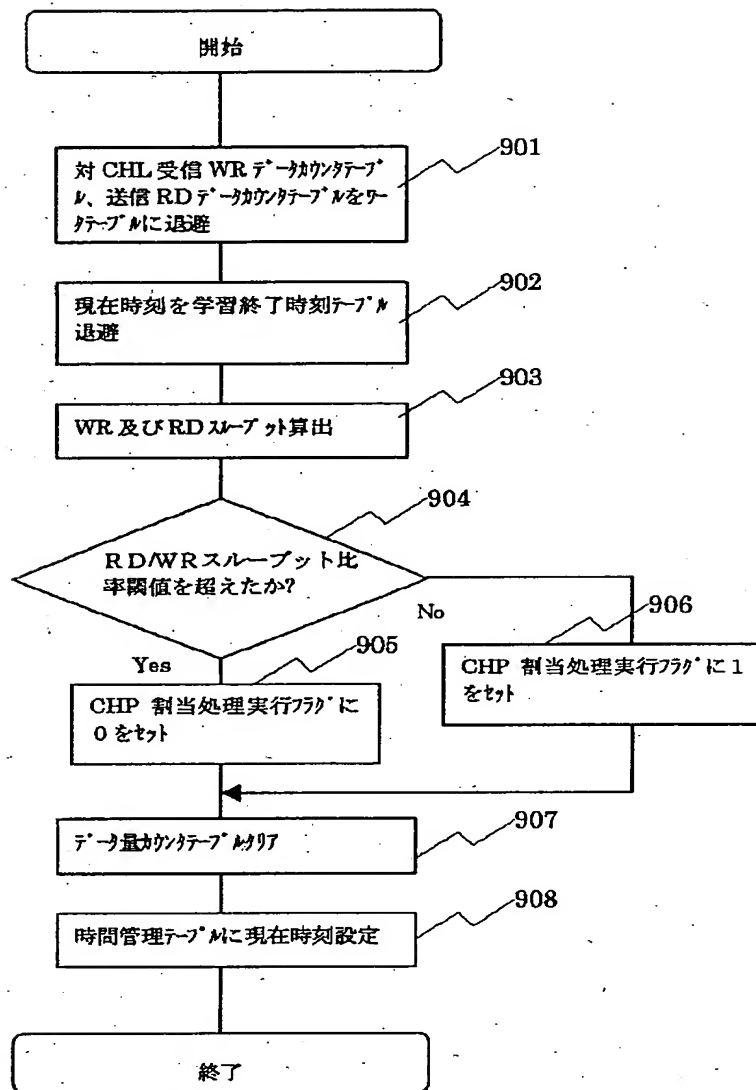
【図7】



【図8】



【図9】



フロントページの続き

(72)発明者 榎 豪紀
 神奈川県小田原市中里322番地2号 株式
 会社日立製作所RAIDシステム事業部内

(72)発明者 塚田 大
 神奈川県小田原市中里322番地2号 株式
 会社日立製作所RAIDシステム事業部内
 Fターム(参考) 5B014 EB04 FB02 GD04 GD34
 5B065 BA01 CA02 CA11 CC08 ZA13

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-222339

(43)Date of publication of application : 11.08.2000

(51)Int.Cl.

G06F 13/14
G06F 3/06
G06F 13/00
G06F 13/10

(21)Application number : 11-024648

(71)Applicant : HITACHI LTD

(22)Date of filing : 02.02.1999

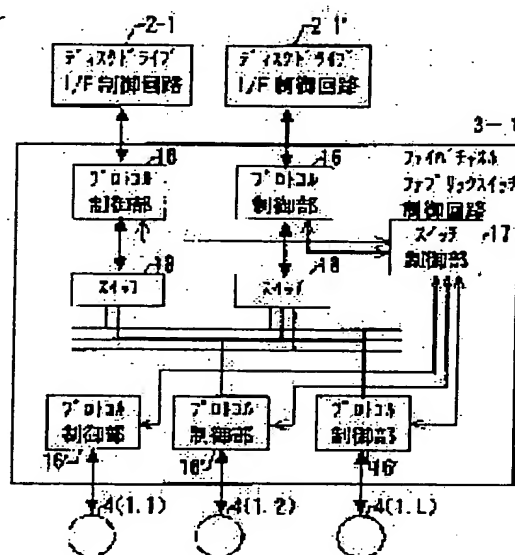
(72)Inventor : ARIGA KAZUHIKA

(54) DISK SUB-SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To connect plural disk drives with a disk drive interface circuit without scarifying transmitting performance by using a fiber channel fabric topology for reducing the number of connection lines by using a fiber channel interface being a serial interface, and for realizing switch connection.

SOLUTION: A fiber channel fabric switch control circuit 3 is provided between a disk drive 4 and a disk drive interface control circuit 2, and protocol control part 16 is provided between a switch 18 in the fabric switch circuit 3 and the disk drive.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2000-222339

(P2000-222339A)

(43)公開日 平成12年8月11日(2000.8.11)

(51)Int.Cl.

識別記号

F I

ターマコード(参考)

G 0 6 F 13/14

3 1 0

G 0 6 F 13/14

3 1 0 F 5 B 0 1 4

3/06

3 0 1

3/06

3 0 1 A 5 B 0 6 5

3 0 5

3 0 5 C 5 B 0 8 3

5 4 0

5 4 0

13/00

3 0 1

13/00

3 0 1 D

審査請求 未請求 請求項の数 5 O L (全 10 頁) 最終頁に続く

(21)出願番号

特願平11-24648

(22)出願日

平成11年2月2日(1999.2.2)

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 有賀 和久

神奈川県小田原市国府津2880番地 株式会

社日立製作所ストレージシステム事業部内

(74)代理人 100068504

弁理士 小川 勝男

Fターム(参考) 5B014 EA02 EA04 EB05 HA09 HA12

5B065 BA01 CA11 CA19 CA30 CE12

EA25 ZA11

5B083 AA08 BB01 BB03 CC02 CD13

EE08 EF11

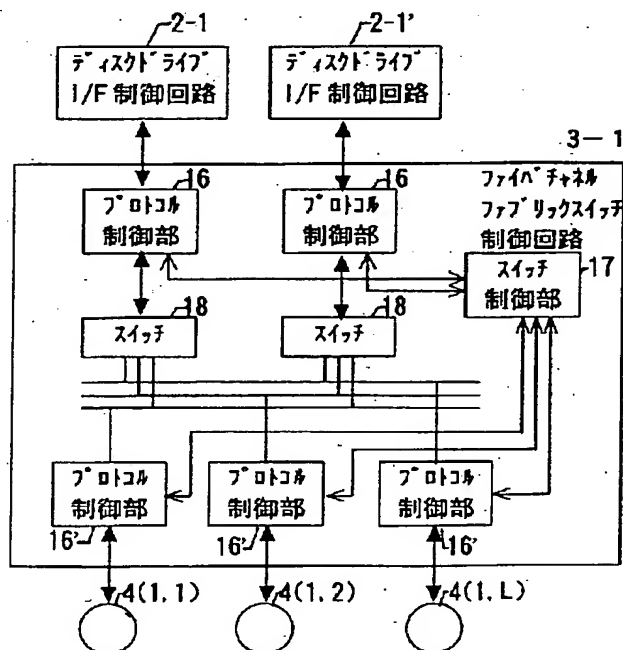
(54)【発明の名称】 ディスクサブシステム

(57)【要約】

【課題】ディスク制御装置とディスクドライブの接続においては、SCSIを用いたインタフェースが主流であるが、ディスクドライブが増加した場合に1本のインタフェースで1対1の接続を行う場合、現状のファイバチャネルを用いたディスクドライブでは、スイッチ接続が出来ない形式となっているので多数のインタフェースが必要となり、実装面で困難が生ずる。

【解決手段】ディスクドライブ4とディスクドライブインタフェース制御回路2との間にファイバチャネル・ファブリック・スイッチ制御回路3を設け、このファイバチャネル・ファブリック・スイッチ回路3内のスイッチ18とディスクドライブとの間にプロトコル制御部16を設ける。

図 3



【特許請求の範囲】

【請求項 1】データを記憶する複数のディスクドライブと、このディスクドライブ及びホストコンピュータからのデータの入出力を制御するディスクアレイ制御部とを有し、このディスクアレイ制御部と前記ディスクドライブとをファイバチャネルにて接続したディスクサブシステムにおいて、

前記ディスクアレイ制御部と前記ディスクドライブとをスイッチ接続したディスクアレイシステム。

【請求項 2】データを記憶する複数のディスクドライブと、このディスクドライブ及びホストコンピュータからのデータの入出力を制御するディスクアレイ制御部とを有するディスクサブシステムにおいて、

前記ディスクドライブと前記ディスクアレイ制御部との間にスイッチとこのスイッチの切換え制御をするスイッチ制御部を設け、前記スイッチと前記ディスクドライブとの間、及び／または前記ディスクアレイ制御部と前記スイッチとの間にプロトコル制御部を設けたディスクサブシステム。

【請求項 3】前記ディスクアレイ制御部と前記スイッチとの間、及び前記スイッチと前記ディスクドライブとの間とはファイバチャネルを用いて接続したものであり、前記スイッチはファイバチャネルファブリックスイッチである請求項 2 に記載のディスクサブシステム。

【請求項 4】ホストコンピュータからのデータの入出力を制御するホストインタフェース制御部と、このホストインタフェース制御部で受けたデータを一時的に格納するキャッシュメモリと、前記データにパリティデータを付加するパリティデータ生成部と、前記データ及び前記パリティデータを記憶する複数のディスクドライブと、このディスクドライブに前記データを書き込むディスクドライブインターフェイスとからなるディスクアレイ制御部とを有するディスクサブシステムにおいて、前記ディスクドライブインターフェイスにプロトコル制御部とスイッチを設け、前記複数のディスクドライブをスイッチ接続したディスクサブシステム。

【請求項 5】ホストコンピュータからのデータの入出力を制御するホストインタフェース制御部と、このホストインタフェース制御部で受けたデータを一時的に格納するキャッシュメモリと、前記データにパリティデータを付加するパリティデータ生成部と、前記データ及び前記パリティデータを記憶する複数のディスクドライブと、このディスクドライブに前記データを書き込むディスクドライブインターフェイスとからなるディスクアレイ制御部とを有するディスクサブシステムにおいて、前記ディスクアレイ制御部と前記ディスクドライブとの間をファイバチャネルを用いて接続し、前記ディスクアレイ部と前記ディスクドライブとの間に、アクセス対象となる前記ディスクドライブの ID 番号検出及びファイバチャネル・プロトコルの制御を行い前記ディスクドラ

イブインタフェースと接続される第一のプロトコル制御部と、各ディスクドライブの ID 番号を記憶しており ID 番号によりスイッチを設定するスイッチ制御部と、前記ディスクドライブにこの ID 番号を割り付ける前記ディスクドライブ 4 と接続される第二のプロトコル制御部とを備えたファブリックスイッチを設けたディスクサブシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ディスクサブシステム、ディスクアレイ、ディスクドライブを内蔵した計算機等の電子機器に関し、特にアレイディスクをファブリックスイッチ接続し高速転送を可能とする技術に関する。

【0002】

【従来の技術】一般に、ディスクアレイにおいてディスク制御装置と複数のディスクドライブとを接続する場合には、特開平 10-171746 号公報に記載のように SCSI インタフェース、若しくはファイバチャネル・アービトレイテッドループ・トポロジが利用されている。

【0003】SCSI インタフェースは、同一線路上にデータを時分割して転送する方式をとっており、イニシエータに対するアクセスは、1 伝送路上に 1 時刻あたり 1 対 1 の通信を行う方式である。

【0004】ファイバチャネル・アービトレイテッドループ・トポロジでは、SCSI インタフェースに対して、シリアルインタフェースによりループ状にイニシエータ、ディスクドライブを接続することができ、フレームに分割されたデータを時分割して転送し、同時に多数デバイスの通信が行え、接続可能ディスクドライブ数も 126 と拡張できる。

【0005】

【発明が解決しようとする課題】今後ディスクドライブの小型化・高密度化により、より多くのディスクドライブを使用することが可能となると考えられる。

【0006】SCSI インタフェースは、1 伝送路上に 1 時刻あたり 1 対 1 の通信を行う方式であるため同時に多数のイニシエータとディスクドライブの通信ができない。また、接続可能なディスクドライブの数も 7 から 15 台と少ない。そのため SCSI を用いたインタフェースでドライブが増加した場合に 1 本のインタフェースで 1 対 1 の接続を行おうとすると、多数のインタフェースが必要となり、実装面で困難が生ずる。また、1 つの制御回路で接続可能なディスクドライブの数が少ないため、多数の制御回路を使用する必要性が生じる。

【0007】一方、ファイバチャネルを使用した場合、ディスクドライブはプロトコルが制御装置とは異なるためにスイッチ接続が出来ず、多数のディスクドライブが同ループを共有するファイバチャネル・アービトレイ

テッドループを用いてループ接続とせざるを得なかった。そのため、同一ループに接続されるディスクドライブ数を増加すると、ディスクドライブのデータ転送速度がループの最大データ転送速度よりも大きくなり、結果的にループの最大データ転送速度以上の効率では転送が行えなくなりSCSIインタフェースと同程度のデータ転送速度でしか接続できなかった。

【0008】

【課題を解決するための手段】上記課題を解決するため本発明では、ディスクドライブと制御装置とをスイッチ接続を可能とするため、プロトコル制御部をファイバチャネル・ファブリック・スイッチとディスクドライブとの間に設ける。

【0009】

【発明の実施の形態】以下、図面を用いて本発明を適用した外部記憶装置（ディスクサブシステム）の実施例を説明する。図1は全体図である。

【0010】図に示す外部記憶装置において、N個のディスクアレイ制御回路（制御部）（1-1）～（1-N）（途中の1-2等は省略、以下同じ）は、上位側はホストコンピュータ（図示せず）に接続され、下位側はM個のディスクドライブインタフェース（ディスクドライブI/F）制御回路（2-1）～（2-M）を備えている。ディスクアレイ制御回路のハード構成の詳細は後述する。M個のファイバチャネル・ファブリック・スイッチ制御回路（3-1）～（3-M）は、ファイバチャネル・インタフェース5によってディスクドライブを制御するディスクドライブインタフェース（I/F）制御回路（2-1）～（2-M）にそれぞれ接続されている。そして一つのファイバチャネル・ファブリック・スイッチ制御回路に対してL個、計M×L個のディスクドライブ（4（1, 1）～4（M, L））は、ファイバチャネル・インタフェース6によって、ファイバチャネル・ファブリック・スイッチ制御回路（3-1）～（3-M）と接続されている。

【0011】また、各ディスクドライブインタフェース制御回路（2-1）～（2-M）、及びデータを格納しておくディスクドライブ4（1, 1）～4（M, L）はそれぞれ個別の識別子（ID番号）をもつ。ファイバチャネル・ファブリック・スイッチ制御回路（3-1）～（3-M）は、ディスクドライブインタフェース制御回路（2-1）～（2-M）から接続するディスクドライブのID番号を受け取り、対応するディスクドライブインタフェース制御回路（2-1）～（2-M）とディスクドライブ4（1, 1）～4（M, L）の1対1の接続を確立する。

【0012】図2にディスクアレイ制御回路（1-1）～（1-N）のハードウェア構成を示す。上位ホストコンピュータ（図示せず）から転送されるデータは、ホストインタフェース制御部7により制御されキャッシュメ

モリ8に一時格納されると共にパリティデータ生成部9によりパリティデータを付加され、データブロックとパリティデータブロックとに分解（全体でM個）される。これらのデータ及びパリティのブロックは、それぞれ対応するインタフェースであるディスクドライブインタフェース制御回路（2-1）～（2-M）によりディスクドライブグループ（図示せず）に格納される。

【0013】上位ホストコンピュータにデータを転送する場合は、転送するデータがキャッシュメモリ8に存在する場合には、そのデータをホストインタフェース制御部7が上位ホストコンピュータに転送する。転送するデータがキャッシュメモリ8に存在しない場合には、ディスクドライブインタフェース制御回路（2-1）～（2-M）がディスクドライブグループより分解されたデータを読み出し、パリティデータ生成部9で分解されたデータを結合した後にキャッシュメモリ8に一時格納するとともにホストインタフェース制御部7が上位ホストコンピュータに転送する。

【0014】なお、以上の例はRAIDを用いた場合のデータ格納方法であり、RAID方式を用いずにデータを格納することも当然可能である。その場合にはパリティデータ生成部9が存在せず上位ホストコンピュータ（図示せず）から転送されるデータは、ホストインタフェース制御部7によりキャッシュメモリ8に一時格納されると共にディスクドライブグループ内の何れかのディスクドライブに格納され、ミラー方式の場合には、複数のディスクドライブに同一のデータを複数格納する。読み出す際にもディスクドライブからデータを読み出し、キャッシュメモリ8に一時格納するとともにホストインタフェース制御部7が上位ホストコンピュータに転送する。

【0015】以下の例もRAIDを使用したディスクサブシステムについて説明するが、RAIDを用いた場合に限らないことはもちろんである。

【0016】図3にファイバチャネル・ファブリック・スイッチ制御回路（3-1）～（3-M）のハードウェア構成を示す。ディスクドライブインタフェース制御回路（2-1）と接続されるプロトコル制御部16（第一のプロトコル制御部）は、アクセス対象となるディスクドライブ4（1, 1）～4（1, L）のID番号検出及びファイバチャネル・プロトコルの制御を行う。ディスクドライブ4（1, 1）～4（1, L）と接続されるプロトコル制御部16'（第二のプロトコル制御部）はディスクドライブ4（1, 1）～4（1, L）にID番号を割り付け、スイッチ制御部17に担当するディスクドライブ4（1, 1）～4（1, L）のID番号を報告する。スイッチ制御部17は、各ディスクドライブ4（1, 1）～4（1, L）のID番号を記憶しており、ディスクドライブインタフェース制御回路（2-1）～（2-M）より受領したID番号によりスイッチ18を

設定し、1対1の接続を確立する。

【0017】尚、プロトコル制御はプロトコル制御部16'側で行うように設定してもよいし、ホストコンピュータからのデータ転送時とホストコンピュータへデータ転送時とや、通常のデータ転送とディスク障害時のデータ移送とでプロトコル制御部16とプロトコル制御部16'とを切り換えるように設定してもよい。

【0018】また、プロトコル制御部16或いはプロトコル制御部16'の何れか一方のみとし、プロトコル制御部16の代わりにID番号検出手段を設ける、或いはプロトコル制御部16'の代わりにID番号割り付け手段を設けてもよい。

【0019】また、ファイバチャネル・ファブリック・スイッチを独立した装置としてではなく、ディスクドライブインターフェイス制御回路(2-1)~(2-M)内にプロトコル制御部とスイッチとを設け、直にディスクドライブ4(1,1)~4(1,L)と接続するようにしてもよい。

【0020】図4にファイバチャネルファブリックスイッチ制御回路(3-1)~(3-M)の動作を示す。

【0021】ディスクアレイ制御回路(1-1)は、M個に分解されたデータをディスクドライブグループ(10-1)に格納する。この際、ディスクアレイ制御回路(1-1)のM個のディスクドライブインタフェース制御回路(2-1)~(2-M)は、ファイバチャネル・ファブリック・スイッチ制御回路(3-1)~(3-M)に対し、ディスクドライブグループ(10-1)に属するディスクドライブのID番号を送信し、スイッチの確立を行う。ファイバチャネル・ファブリック・スイッチ制御回路(3-1)~(3-M)内のプロトコル制御部16(図3参照)は、ID番号を検出し、スイッチ制御部17にスイッチ接続の切替を要求する。そしてディスクドライブに合わせたプロトコル制御を行う。スイッチ制御部17(図3参照)はスイッチ18(図3参照)を接続要求もとのディスクアレイ制御回路(1-1)と接続要求先のディスクドライブグループ(10-1)に属するディスクドライブ4とを接続するよう切り替える。

【0022】このとき、ディスクアレイ制御回路(1-1)は、ディスクドライブグループ(10-1)とファイバチャネル・ファブリック・スイッチ制御回路(3-1)~(3-M)を介して1対1で対応しているので、他のディスクアレイ制御回路(1-N)と他のディスクドライブグループ(10-2)は独立して他のデータ転送を行うことが出来る。つまり、ディスクアレイ制御回路(1-N)がディスクドライブグループ(10-L)に対する接続の確立を行っても、ディスクアレイ制御回路(1-1)とディスクドライブグループ(10-1)及びディスクアレイ制御回路(1-N)とディスクドライブグループ(10-L)との接続は互いに独立して動

作することができるので、それぞれのディスクアレイ制御回路及びディスクドライブ間で可能となる最高のデータ転送速度でデータ転送を行うことができる。

【0023】尚、詳細は説明しないが、スイッチ制御部17は上記のスイッチ切換えを行うと共に、データ読み書きの際にディスクドライブが既に読み書きを出来る状態になったという信号を受けてスイッチ18の接続切換えを行うことで転送時間を有効に最大限確保することができる。

【0024】図5に本発明の拡張された実施例を示す。先に示した実施例において、ファイバチャネル・ファブリック・スイッチ制御回路3のプロトコル制御部16とディスクドライブ4とを1対1で対応させて接続していた部分を、プロトコル制御部16からファイバチャネル・アービトレイテッド・ループ制御回路11を介して複数のディスクドライブ4をループ接続するしている。この様に接続することで、安価なディスクドライブ4を多数の接続することで大容量のディスクドライブを備えた場合と同等な性能にできる。この場合でも、全てのディスクドライブがループ接続となる訳ではなく、見かけ上はファイバチャネル・アービトレイテッド・ループ制御回路11と多数のディスクドライブ4で一つのディスクドライブ4であるので、アクセス性能は低下することがない。

【0025】また図示はしないが、ディスクドライブのアクセス速度に対し、ファイバチャネルインタフェースの最大データ転送速度に充分余裕がある場合には、複数のディスクドライブ4をファイバチャネル・アービトレイテッド・ループ制御回路11に接続し、複数のディスクドライブを同一ループ内に接続し、ファイバチャネルの最大転送レートを複数のディスクドライブ4で共有することで、アクセス性能を低下させることなくディスクドライブ4を増加させることも可能である。

【0026】図6に図5に示した実施例に用いるアービトレイテッドループ制御回路11のハードウェア構成図を示す。

【0027】アービトレイテッドループ制御回路11は、ループバイパス回路13と複数のディスクドライブ接続ポート12、及びファブリックスイッチ接続ポート15からなる。ディスクドライブ4からはループバイパス回路切替信号14が出力され、ディスクドライブ障害時にはポートをバイパスさせ、ループを切断することなく、他の動作しているディスクドライブへ影響を与えずにディスクドライブの取り外し、追加を行うことを可能とする。

【0028】図7に本発明の他の拡張された実施例を示す。

【0029】本実施例は、各ファイバチャネル・ファブリック・スイッチ制御回路(3-1)~(3-M)に接続されるスペアディスク制御回路19と、このスペアデ

ィスク制御回路19に接続される複数のスペアディスクドライブ(4-a)、(4-b)を備えている。そしてファイバチャネル・ファブリック・スイッチ制御回路3内では、故障したディスクドライブ4を含むディスクドライブグループ(図では4(1, 2)のディスクドライブグループ)と接続しているプロトコル制御部16'

(図3参照)は、スイッチ18を介してスペアディスク制御回路19と接続しているプロトコル制御部16'に接続される。何れかのディスクドライブが障害を起こした場合、ディスクアレイ制御回路(1-1)~(1-N)は、スペアディスクドライブ(4-a)または(4-b)にデータの再構築を行う。

【0030】特定のディスクドライブ4にエラーが多発し故障のおそれが出た場合には、エラーが多発するディスクドライブ4のデータをスペアディスクドライブ(4-a)または(4-b)に移管させ再構築を行う。ディスクドライブ4が完全に破損してしまいデータの移管が不可能な場合には、破損したディスクドライブ4のディスクドライブグループのデータを用いて、図2に示したキャッシュメモリ8とパリティデータ生成部9にて破損データを再生しスペアディスクドライブ(4-a)または(4-b)に書き込む。

【0031】或いは、スペアディスク制御回路19が独立して行うようにしてもよい。そのためこのスペアディスク制御回路19内にキャッシュメモリやパリティデータ再生部を備える。そして、ディスクドライブ4が完全に破損した場合には、残りのディスクドライブグループのデータをスペアディスク制御回路19で読み込み、破損データを再生してスペアディスクドライブ(4-a)または(4-b)に書き込むようにする。

【0032】そのため、故障したディスクドライブ4或いは故障箇所を修復するためにパリティデータを含め分割されたデータを記憶した各ディスクドライブからスペアディスク制御回路19へのアクセスと、ディスクアレイ制御回路(1-1)~(1-N)を介して行うディスクドライブ4(図では4(1, 1)及び4(1, L)のディスクドライブグループ)とホストコンピュータからのデータアクセスとが独立して動作可能となることで、ホストコンピュータのデータアクセスに影響を与えずにデータの再構築を行うことを可能とする。

【0033】また、障害ディスクドライブが正常ディスクドライブと取り替えられた場合も同様にして、スペアディスク制御回路15がファイバチャネル・ファブリック・スイッチ制御回路(3-1)~(3-M)に対し、スペアディスクドライブ(4-a)、(4-b)と、障害ディスクドライブから取り替えられた正常ディスク

ドライブと1対1の接続を行い、ディスクアレイ制御回路(1-1)~(1-N)とディスクドライブグループ(10-1)~(10-L)(図4参照)とのアクセスを妨げることなく、独立してデータのコピーを行うことで、ホストコンピュータからのアクセスにまったく影響なく障害ディスクドライブの復旧を行うことができる。

【0034】

【発明の効果】本発明により、シリアルインタフェースであるファイバチャネルインタフェースを用い接続線数を減少させ、さらにスイッチ接続を可能とするファイバチャネル・ファブリック・トポロジを用いることでディスクドライブインタフェース回路に多数ディスクドライブを伝送性能を犠牲にすることなく接続することが可能となる。また、各制御装置、ディスクドライブグループ毎に接続を動的に切り替えることで、少数のディスクドライブ制御回路で多数のディスクドライブを制御することができる。更に、ディスクドライブ障害時のデータ移行をディスクドライブインタフェース制御回路とディスクドライブのデータ転送と独立して行うことでシステムの信頼性を向上させることができる。

【図面の簡単な説明】

【図1】本発明実施例の全体図である。

【図2】ディスクアレイ制御回路の詳細図である。

【図3】ファイバチャネルファブリックスイッチ制御回路の詳細図である。

【図4】ファイバチャネルファブリックスイッチの接続図である。

【図5】ファイバチャネルファブリックスイッチとアービトレイテッドループの接続図である。

【図6】ファイバチャネルアービトレイテッドループ制御回路の詳細図である。

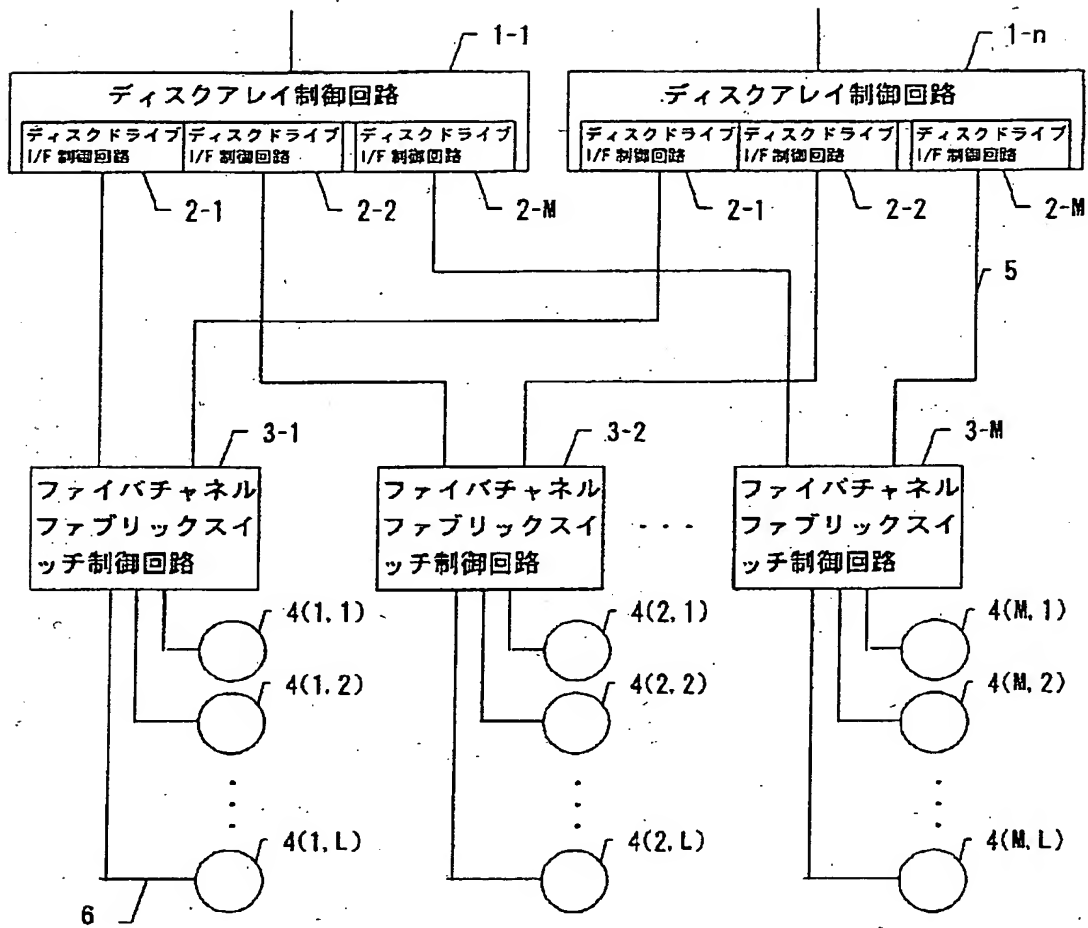
【図7】スペアディスク制御回路の接続図である。

【符号の説明】

1…ディスクアレイ制御回路、2…ディスクドライブインタフェース制御回路、3…ファイバチャネルファブリックスイッチ制御回路、4…ディスクドライブ、5…ファイバチャネルインタフェース、6…ファイバチャネルインタフェース、7…ホストインタフェース制御部、8…キャッシュメモリ、9…パリティデータ生成部、10…ディスクドライブグループ、11…ファイバチャネルアービトレイテッドループ制御回路、12…ディスクドライブ接続ポート、13…ループバイパス回路、14…ループバイパス信号切替信号、15…ファブリックスイッチ接続ポート、16…プロトコル制御部、17…スイッチ制御部、18…スイッチ、19…スペアディスク制御回路。

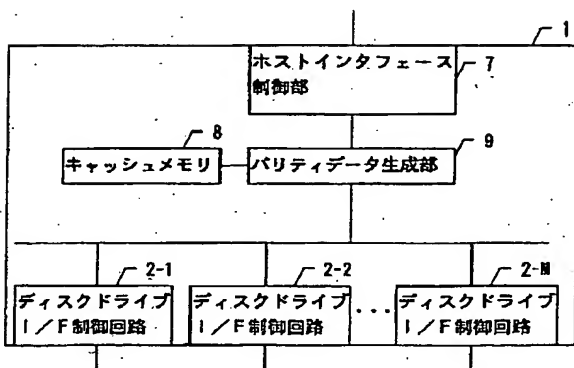
【図1】

図 1



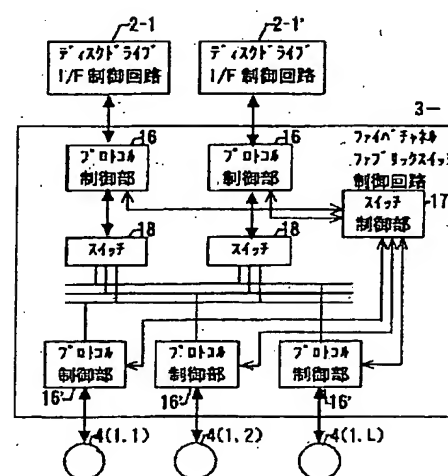
【図2】

図 2



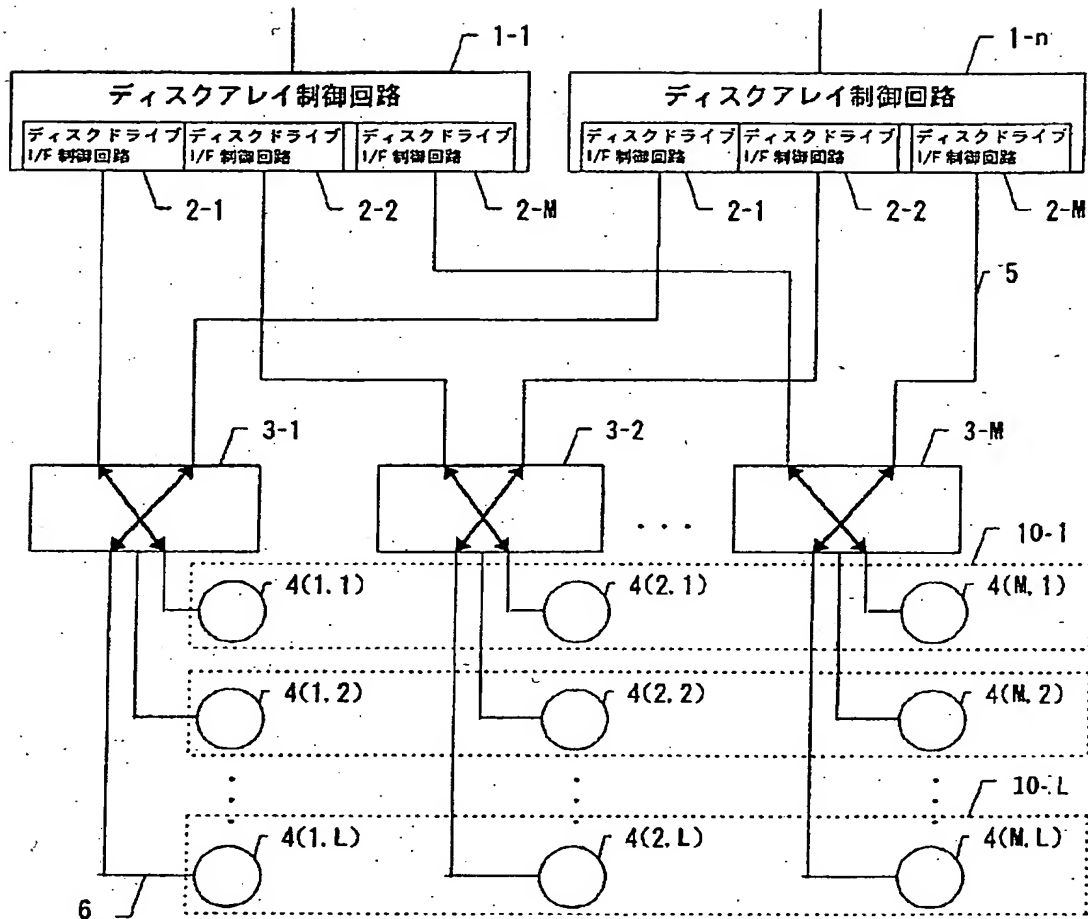
【図3】

図 3



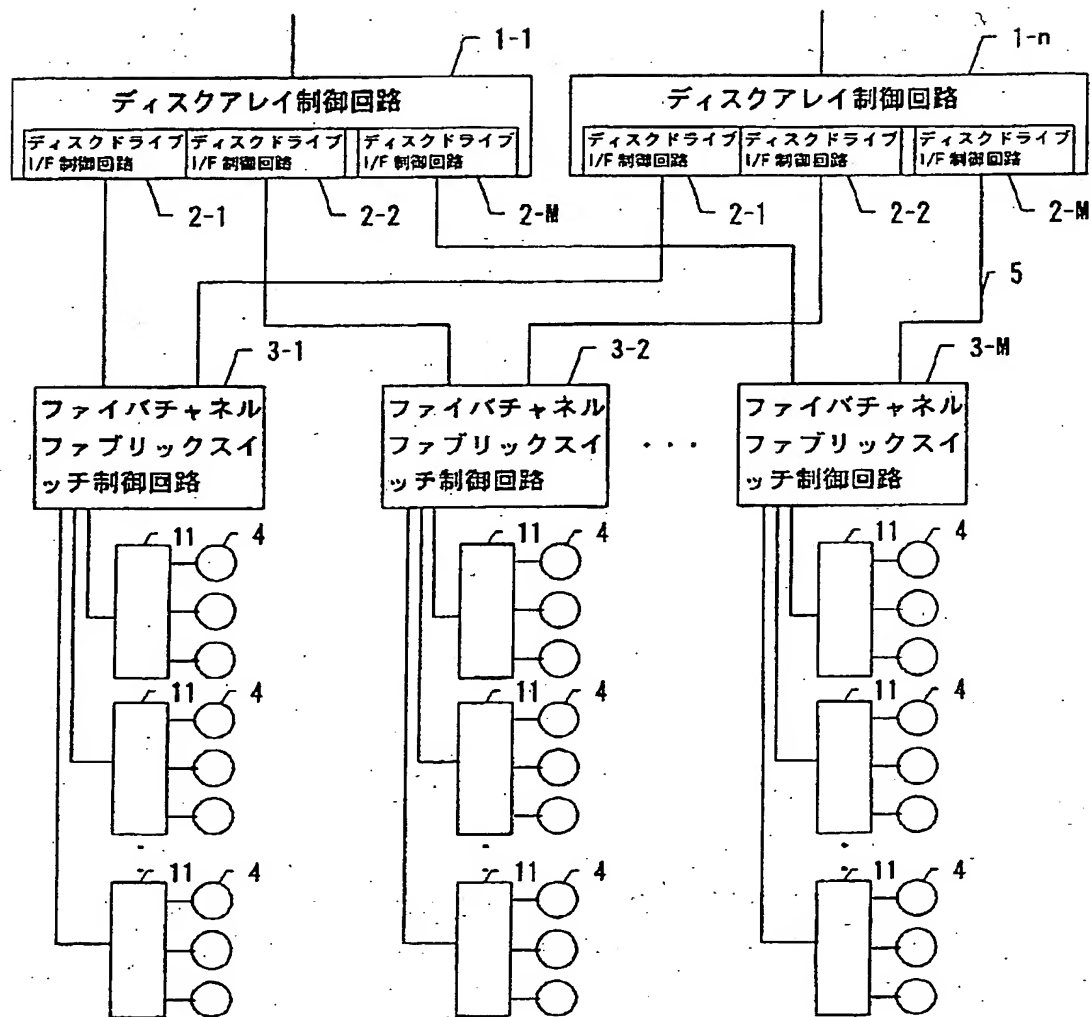
【図4】

図4



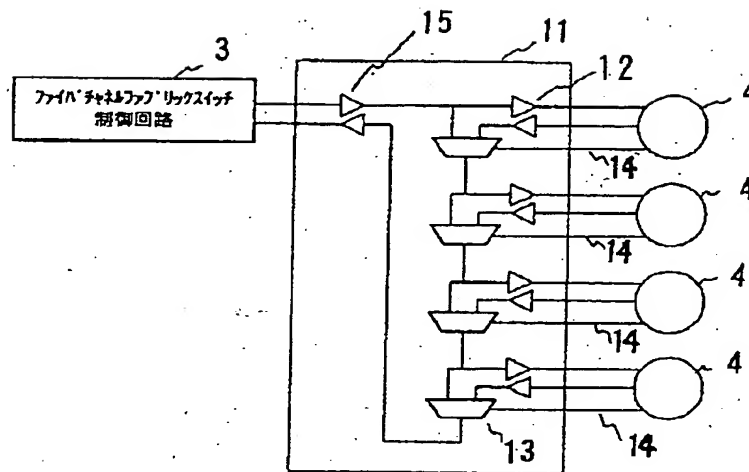
【図5】

図5



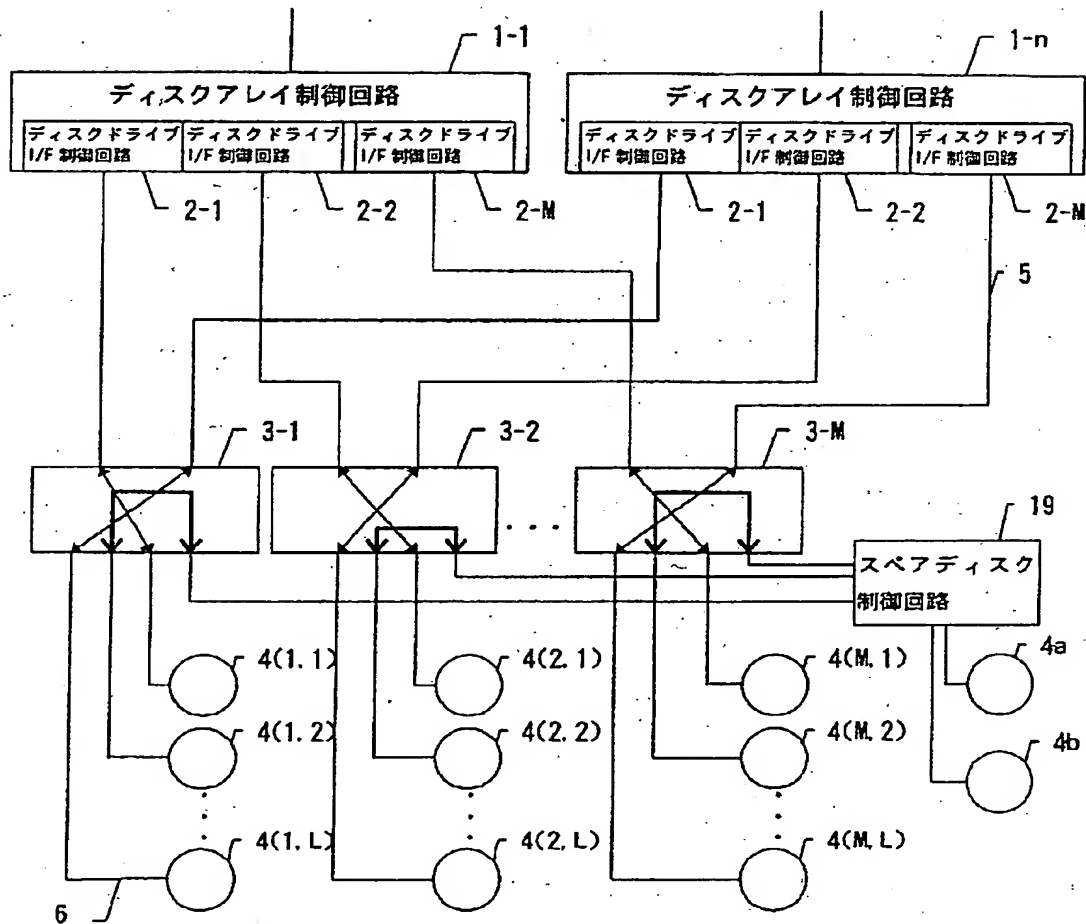
【図6】

図6



【図 7】

図 7



フロントページの続き

(51) Int. Cl.⁷

G 0 6 F 13/10

識別記号

3 4 0

F I

G 0 6 F 13/10

テーマコード (参考)

3 4 0 A

[Patent Document]

1. Japanese Patent Laid Open

No. 2003-303055

Disk storage system having disk arrays connected with disk adapters through switches

Hitachi, Ltd.

Inventor(s): Tanaka, Katsuya ; Fujimoto, Kazuhisa

Application No. 10/212882, Filed 20020807, A1 Published 20031009

Abstract:

A disk storage system has high throughput between a disk adapter of a disk controller and a disk array. The disk adapter of the disk controller is connected to the disk array through switches. Data on a channel between the switch and a RAID group is multiplexed in the switch to be transferred onto a channel between the switch and the disk adapter and data on the channel between the switch and the disk adapter is demultiplexed in the switch to be transferred onto the channel between the switch and the RAID group. A data transfer rate on the channel between the disk adapter and the switch is made higher than that on the channel.

US.Class: 711114 711154

(19) 日本国特許庁 (J P) (12) 公開特許公報 (A)

(51) Int. Cl.
G 06 F 3/06 3/01 3/02 5/40

(11) 特許出願公開番号
特開 2003-303055
(P 2003-303055A)
(43) 公開日 平成 15 年 10 月 24 日 (2003.10.24)

(52) 出願番号 特願 2002-106262 (P 2002-106262)
(22) 出願日 平成 14 年 4 月 9 日 (2002.4.9)

審査請求 未請求 請求項の数 9 OL (全 14 頁)

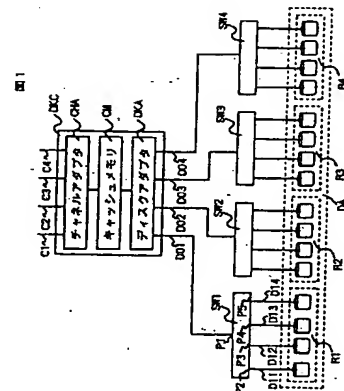
(71) 出願人 000005108
株式会社日立製作所
東京都千代田区神田墨田区四丁目 6 番地
(72) 発明者 田中 勝也
東京都国分寺市東葛ヶ丘一丁目 280 番地
株式会社日立製作所中央研究所内
(72) 発明者 藤本 和久
東京都国分寺市東葛ヶ丘一丁目 280 番地
株式会社日立製作所中央研究所内
(74) 代理人 100099288
弁理士 伊藤 修 (外 1 名)
F ターム (参考) 5B05 B401 C004 C007 CA12 CA15
CA30 C008 CE11 CH01

(54) 【発明の名称】 ディスクアダプタとディスクアレイをスイッチを介して接続したディスク装置

(57) 【要約】

【課題】 ディスクコントローラのディスクアダプタとディスクアレイ間のスループットが高いディスク装置を提供することにある。

【解決手段】 ディスクコントローラ (0X) のディスクアダプタ (0XA) とディスクアレイ (0A) をスイッチ (S1), S1 2, S13, S14 を介して接続する。スイッチ (S1) と R A I D グループ (R1) 間のチャネル (D11, D12, D13, D14) 上のデータをスイッチ (S1) において多重化してスイッチ (S1) とディスクアダプタ (0XA) 間のチャネル (D01) に転送し、スイッチ (S1) とディスクアダプタ (0XA) 間のチャネル (D01) 上のデータをスイッチ (S1) において逆多重化してスイッチ (S1) と R A I D グループ (R1) 間のチャネル (D11, D12, D13, D14) に転送する。ディスクアダプタ (0XA) とスイッチ (S1) 間のチャネル (D01) 上のデータ転送速度を、チャネル (D11, D12, D13, D14) のデータ転送速度より高くする。



【特許請求の範囲】

【請求項 1】 ディスクコントローラとディスクアレイからなり、前記ディスクコントローラはチャネルアダプタとキャッシュメモリとディスクアダプタを有するディスク装置において、

前記ディスクアダプタと前記ディスクアレイを、バッファメモリを有するスイッチを介して接続し、前記スイッチは、前記ディスクアダプタが接続されたポートと前記ディスクアレイを構成するディスクドライバが接続された各ポートとの間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信優先情報にしたがって行うことを特徴とするディスク装置。

【請求項 2】 ディスクコントローラと複数のディスクアレイからなり、前記ディスクコントローラはチャネルアダプタとキャッシュメモリとディスクアダプタを有するディスク装置において、

前記ディスクアレイはループ状に接続した複数のディスクドライバからなり、前記ディスクアダプタと前記複数のディスクアレイとをバッファメモリを有するスイッチを介して接続し、

前記ディスクアダプタと前記スイッチ間のチャネル当りデータ転送速度を、前記スイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

前記スイッチは、前記ディスクアダプタが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信優先情報にしたがって行うことを特徴とするディスク装置。

【請求項 3】 ディスクコントローラとディスクアレイからなり、前記ディスクコントローラはチャネルアダプタとキャッシュメモリとディスクアダプタを有するディスク装置において、

前記ディスクアダプタと前記ディスクアレイを、バッファメモリを有するスイッチを介して接続し、同一のスイッチに接続したディスクドライバの組み合わせで R A I D グループを構成し、

前記ディスクアダプタと前記スイッチ間のチャネル当りデータ転送速度を、前記スイッチと前記ディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

前記スイッチは、前記ディスクアダプタが接続されたポートと前記 R A I D グループを構成するディスクドライバが接続された各ポートとの間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信優先情報にしたがって行うことを特徴とするディスク装置。

【請求項 4】 第 1 のディスクコントローラと第 2 のディスクコントローラと複数のディスクアレイからなり、第 1 のディスクコントローラは第 1 のチャネルアダプタ

と第 1 のキャッシュメモリと第 1 のディスクアダプタを有し、

第 2 のディスクコントローラは第 2 のチャネルアダプタと第 2 のキャッシュメモリと第 2 のディスクアダプタを有するディスク装置において、

第 1 のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第 1 のスイッチを介して接続し、且つ第 2 のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第 2 のスイッチを介して接続し、さらに第 1 のスイッチと第 2 のディスクアダプタを接続し、第 2 のスイッチと第 1 のディスクアダプタを接続し、

第 1 のディスクアダプタと第 1 のスイッチ間、および第 2 のディスクアダプタと第 1 のスイッチ間のチャネル当りデータ転送速度を第 1 のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

第 2 のディスクアダプタと第 2 のスイッチ間、および第 1 のディスクアダプタと第 2 のスイッチ間のチャネル当りデータ転送速度を第 2 のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

第 1 のスイッチは、第 1 のディスクアダプタまたは第 2 のディスクアダプタが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信優先情報にしたがって行うことを特徴とするディスク装置。

【請求項 5】 第 1 のディスクコントローラと第 2 のディスクコントローラと複数のディスクアレイからなり、第 1 のディスクコントローラは第 1 のチャネルアダプタと第 1 のキャッシュメモリと第 1 のディスクアダプタを有し、

第 2 のディスクコントローラは第 2 のチャネルアダプタと第 2 のキャッシュメモリと第 2 のディスクアダプタを有するディスク装置において、

第 1 のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第 1 のスイッチを介して接続し、且つ第 2 のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第 2 のスイッチを介して接続し、さらに第 1 のスイッチと第 2 のディスクアダプタを接続し、第 2 のスイッチと第 1 のディスクアダプタを接続し、

第 1 のディスクアダプタと第 1 のスイッチ間、および第 2 のディスクアダプタと第 1 のスイッチ間のチャネル当りデータ転送速度を第 1 のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

第 2 のディスクアダプタと第 2 のスイッチ間のチャネル当りデータ転送速度を第 2 のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

りデータ転送速度を第1のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

第2のディスクアダプタと第2のスリット間、および第1のディスクアダプタと第2のスリット間のチャネル当りデータ転送速度を第2のスリットと前記複数のディスクアダプタ間のチャネル当りデータ転送速度より高く設定し、

第1のスイッチと第2のスイッチを、第1のディスクアダプタと第2のディスクアダプタを接続したチャネルと同等のデータ転送速度を有するチャネルと、第2のディスクアダプタと第1のディスクアダプタを接続したチャネルと同等のデータ転送速度を有するチャネルと、を介して接続し、第1のスイッチは、第1のディスクアダプタまたは第2のディスクアダプタまたは第2のスイッチが接続されたポートと前記対象のディスクアダプタが接続された各ポートとの間で、ポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行い、

第2のスイッチは、第1のディスクアダプタまたは第2のディスクアダプタまたは第1のスイッチが接続されたポートと前記複数のディスクアレイが接続された各ポートとの間でポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行うことを特徴とするディスク装置。

【請求項6】請求項1乃至請求項5のいずれかの請求項記載のディスプレイ装置において、

前記ディスプレイからのデータ読み出し時には、前記ディスプレイから前記スイッチに転送されるデータを前記スイッチにおいて多重化して前記ディスプレイに転送し、

前記ディस्कアレイへのデータ書き込み時には、前記ディスクアダプタから前記スイッチに転送されるデータを前記前記スイッチにおいて逆多重化して前記ディスクアレイに転送することを特徴とするディस्क装置。

ディस्कアレイへのデータ書き込み時に、前記ディスクアダプタは、前記ポート間の接続の切り替えが個別的に行われるように、送出するフレームに送出先情報を設定し、ディスクアレイからディスクアダプタへのデータ読み出し時に、前記スイッチは、ラウンドロビン方式により前記ポート間の接続を切り替えることを特徴とするディスク装置。

【請求項8】 請求項7記載のディスク装置において、周期的に切り替えるポート数を、ディスクアダプタとスライイチツチ間のチャネル当りデータ転送速度の、スライイチとディスクアレイ間のチャネル当りデータ転送速度に對する比、と同程度に設定することを特徴とするディスク装

【請求項9】 請求項1乃至請求項5のいずれかの請求項記載のディスク装置において、前記ディスクアダプタと前記スライツ間を光ファイバケーブルで接続し、前記スライツと前記ディスクアレイ間をメタルケーブルで接続することを特徴とするディスク装置。

【発明の詳細な説明】

【0001】
【発明の属する技術分野】本発明は、コンピュータシステムにおける2次記憶装置に関し、特に入出力データ転送性能が高いディスク装置に関する。

【0002】現在の技術（コンピュータシステムにおいて）は、CPU（中央処理装置）が必要とするデータは2次元配列装置に保存され、CPUなどが必要とするときに必要に応じて2次元配列装置に対してデータの書き込みおよび読み出しを行う。このため配列装置としては、一般に不揮発的な配列媒体が使用され、代表的なものとして磁気ディスク装置や、光ディスクなどのディスク装置がある。近年、高度情報化に伴い、コンピュータシステムにおいて、この種の2次元配列装置の高性能化が要求されている。

【0003】図9に、従来のディスク装置のブロック図を示す。図9において、ディスク装置はディスクコントローラDKCとディスクアレイDAで構成される。ディスクコントローラDKCは、上位側CPU（図示せず）

と、ディस्क装置に接続するチャネルアダプタCHA、と、ディスクアレイDAに対して読み書きするデータを一時的に保存するキャッシュメモリCMと、ディスクコントローラDKCと、ディスクアレイDAを接続するディスタダブダブDKAとなる。チャネルアダプタCHAとキャッシュメモリCMとディスクアダブダブDKAは、バスまたはスイッチで相互接続されている。チャネルアダプタCHAはC1、C2、C3、C4の4本のチャネルでCPUと接続している。ディスクアダブダブDKAはD1、D2、D3、D4の4本のチャネルでディスクアレイと接続している。ここでディスクアレイDAはディスクグループR1、R2、R3、R4からなる。ディスクアレイDAにおいてRAIDシステムを構成する場合は、R1、R2、R3、R4がそれぞれRAIDグループを構成する。

【0004】チャネルC1、C2、C3、C4から入力された書き込みデータは、キャッシュメモリCMにデータを書き込むと同時に、格データをブロックサイズ単位で分割し、チャネルD1、D2、D3、D4の内3チャネルにはブロック単位に分割されたデータを、残りのチャネルには前記分割データから計算したパリティを、ディスクアダプタDKAからディスクアレイDAへ送信する。データ読み出し時は、まずキャッシュメモリCMに格当データの有無を調べる。有る場合は、キャッシュメモリCMからディスクアレイDAへ送信する。

ユメメモリCMからチャネルアダプタCHAを介してキヤッシュメモリ内蔵出しデータをCPUへ送信する。キヤッシュメモリCM内無き場合にはディスクアダプタDKAは、D1、D2、D3、D4を介してディスクアレイDAからブロック単位に分割されたデータを読み出し、チャネルアダプタCHAを介して読み出しデータをCPUへ送信する。この種の従来技術を図1の従来技術と呼ぶ。第1の従来技術として関連するディスク装置は、例えば、日産B社刊の「日産コンピュータ別冊メインフレーム'98」(1998年)第144頁から第153頁に記載されているディスク装置がある。

【0005】 ディスクアダプタとディスクアレイを、スライツを介して接続したディスク装置が、特開平5-173732号の「マルチチャネルデータおよびバリエーションの交換デバイス」に開示されている。以下、該公報に記載の従来の技術を第2の従来技術と呼ぶ。第2の従来技術によれば、ディスクアレイに関連したバス本数と、ディスクアダプタに関連したバス本数とを独立に設定できる。ディスクアダプタとディスクアレイを、バックアップ6-19627号の「画像記憶装置」に開示されている。以下、該公報に記載の従来の技術を第3の従来技術と呼ぶ。第3の従来技術によれば、ディスクアダプタとディスクアレイの従来技術とは異なり、ディスクアダプタとディスクアレイの間のデータ伝送速度を任意に設定でき、ディスクの回転待ち時間の影響を低減できる。

【0006】
【発明が解決しようとする課題】 ネットワーク技術の進歩に伴い、例えば当りのデータ転送速度は年々増加している。例えばディスク装置に使用されるフアイバチネルでは、現状でチャネル当りのデータ転送速度が1 Gbpsから2 Gbpsであるが、近い将来4 Gbpsから10 Gbpsへ高速度化されることが予定されている。CPUとチャネルアダプタ間（以下フロントエンドと呼ぶ）のスループットはこの高速度化に従うことが予想される。ところが、ディスクアダプタとディスクアレイ（以下バックエンドと呼ぶ）のスループットは以下の理由により、フロントエンドほど高速度化されないとして予想される。第1の理由は、ディスクドライブは機械部品を含むので、電子、光素子のみ高速度化を行えば良いフロントエンドに比べ、高速度化が難しいこと、である。第2の理由では、たとえディスクドライブが高速度化したとしても、全てのディスクドライブ毎に高速インターフェイスを搭載するのは、多数のディスクドライブを有するディスク装置の高コスト化を招くことである。

【0007】第1の従来技術では、チャネルアダプタのチャネル当りのデータ転送速度を向上させても、フロンテントとバックエンドのスループット輻輳により、デイスク装置の性能が向上しないという問題があった。また、バックエンドのスループット向上のために従来ポートを多数、デイスクアダプタに設けることも考えられる

が、ディスクアダプタのポート数増加は制御を複雑とする。第2の従来技術では、ディスクアダプタとディスクアレイとの間にスイッチを適用することができるが、チャネル当り増設ポート数を増加させることができるが、チャネル当りのデータ転送速度はディスクアレイのデータ転送速度に制限されるので、ディスクアダプタとディスクアレイ間のスループットが性能ネックになるという問題があった。第3の従来技術は、ディスクの回転待ち時間の影響を低減できる技術であり、フロントエンドとバックエンドのスループット平衡は低減できないという問題があった。

【0008】本発明の目的は、ディスクアダプタとディスクスレーブ間のスループットが高いディスク装置を提供することにある。本発明の他の目的は、ディスクアダプタとディスクスレーブ間のスループットが高く、且つディスタンスとドライブ接合数が多いディスク装置を提供することにある。本発明のさらに他の目的は、信頼性が高いディスクスレーブを有するディスク装置を提供することである。本発明のさらに他の目的は、信頼性が高いディスクアダプタとディスクスレーブ間ネットワークを有するディスク装置を提供することにある。本発明のさらに他の目的は、信頼性およびスループットが高いディスクアダプタとディスクスレーブ間ネットワークを有するディスク装置を提供することにある。本発明のさらに他の目的は、ディスクからの読み出しおよびディスクへの書き込みをスループット化できるディスク装置を提供することにある。本発明のさらに他の目的は、高スループットを維持できるディスク装置を提供することである。本発明のさらに他の目的は、高スループットで低コストなディスク装置を提供することである。

【0009】
 【課題】解決するための手段】上記目的を達成するため、本発明は、ディスクコントローラとディスクアレイとを、例えば、ディスクコントローラはチャネルアダプタとキャパシタメモリモジュールとディスクアダプタを有するディスク装置であり、ディスクアダプタとディスクアレイを、バスマスタメモリモジュールを有するスリッチを介して接続し、ディスクアダプタとスリッチ間のチャネル当りデータ転送速度を、スリッチとディスクアレイ間のチャネル当りデータ転送速度より高く設定し、ディスクアレイを構成するディスクが接続されたポートと、スリッチは、ディスクアダプタが接続されたポートを、入力されたフレーム毎に、被フレーム内の送信先情報にしたがって行っている。また、前記ディスクアレイはループ状に接続した複数のディスクアダプタからなり、前記ディスクアダプタと前記複数のディスクアダプタとをバスマスタメモリモジュールを有するスリッチを介して接続し、ディスクアダプタとスリッチ間のチャネル当りデータ転送速度を、スリッチと複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、

スイッチは、ディスクアダプタが接続されたポートと複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行っている。また、前記ディスクアダプタと前記ディスクアレイを、バッファメモリを有するスイッチを介して接続し、同一のスイッチに接続したディスクドライブの組み合わせで、RAIDグループを構成し、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度を、スイッチとディスクアレイ間のチャネル当りデータ転送速度より高く設定し、スイッチは、ディスクアダプタが接続されたポートとRAIDグループを構成するディスクドライブが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行っている。また、第1のディスクコントローラと第2のディスクコントローラと複数のディスクアレイからなり、第1のディスクコントローラは第1のチャネルアダプタと第1のキャッシュメモリと第1のディスクアダプタを有し、第2のディスクコントローラは第2のチャネルアダプタと第2のキャッシュメモリと第2のディスクアダプタを有するディスク装置であり、第1のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第1のスイッチを介して接続し、且つ第2のディスクアダプタと前記複数のディスクアレイとをバッファメモリを有する第2のスイッチを介して接続し、さらに第1のスイッチと第2のディスクアダプタを接続し、第2のスイッチと第1のディスクアダプタを接続し、第2のディスクアダプタと第2のスイッチの間、および第1のディスクアダプタと第2のスイッチ間のチャネル当りデータ転送速度を第2のスイッチと前記複数のディスクアレイ間のチャネル当りデータ転送速度より高く設定し、第1のスイッチは、第1のディスクアダプタまたはポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行っている。また、第2のスイッチは、第1のディスクアダプタまたはポートと前記複数のディスクアレイが接続された各ポートとの間でのポート間の接続の切り換えを、入力されたフレーム毎に、該フレーム内の送信先情報にしたがって行っている。また、さらに、上記第1のスイッチと第2のスイッチを、上記第1のディスクアダプタと第2のスイッチ間を接続したチャネルと同等のデータ転送速度を有するチャネルと、第2のディスクアダプタと第1のスイッチ間を接続したチャネルと同等のデータ転送速度を有するチャネルと、を介して接続している。また、前記ディスクアレイからのデータ読み出し時には、前記ディスクアレイから前記スイッチに転送されるデータを前記スイッチに介して多量化して前記ディスクアダプタに転送し、前記ディ

IDシステムを構築する場合は、ディスクグループR 1、R2、R3、R4を、それぞれRAIDグループとする。本実施の形態では、4個のディスクドライブでRAIDグループを構成しているが、RAIDグループを構成するドライブ数を4個に限るものではない。各ディスクグループへのデータ読み出しまたは書き込み時のデータの流れを、ディスクグループR1を例にして述べる。ここでR1はRAIDレベル5のRAIDグループである。チャネルC1、C2、C3、C4からディスクグループR1へ書き込むためにCPUから送信されたデータは、ディスクアダプタDKAにおいてブロック単位に分割されると同時に、該ブロック単位に分割されたデータからパリティが生成される。該ブロック単位に分割されたデータと、生成されたパリティは、チャネルD01を通じてスイッチSW1へ入力される。スイッチSW1は、RAID制御に伴い、該ブロック単位に分割されたデータと、生成されたパリティとをルーティングし、チャネルD11、D12、D13、D14へ分配する。データ読み出し時は、ディスクアダプタDKAは、D11、D12、D13、D14を介してディスクグループR1からブロック単位に分割されたデータを読み出し、スイッチSW1でシリアル化して、チャネルD01を介して読み出しデータを受信する。

[0013] 図9に示した従来のディスク装置では、ディスクアダプタDKAに接続したチャネルD1、D2、D3、D4上で、既にディスクアレイへの書き込みデータおよびパリティが別々のチャネルに分配されていた。それに対し、本実施の形態のディスク装置においては、スイッチSW1通過後に別々のチャネルに分配される点が従来と異なる。

[0014] 次に、本実施の形態のディスク装置の特徴であるスイッチの動作を、スイッチSW1を例にとり説明する。SW2～SW4の動作もSW1の動作と同様である。図1に示すように、スイッチSW1は入力ポートP1、P2、P3、P4、P5を有する。ポートP1、P2、P3、P4、P5は、全二重通信可能な出力ポートであり、ポート毎にバッファメモリを有している。スイッチSW1の内部構成を図2と図3に示す。簡単に説明する。また、チャネルD01、D11、D12、D13、D14上を流れるデータは、フレーム単位で送信され、かつデータは8B10B変換で符号化されている。

[0015] 図2は、ポートP1からブロック内のフレームを入力し、ポートP2、P3、P4、P5から出力する場合を示す。これはディスクアレイへの書き込み時のスイッチ動作に相当する。スイッチSW1は図2に示すように、クロスバスイッチSWと、スイッチコントローラCTLからなる。クロスバスイッチSWは5×5のクロスバスイッチであり、入力ポートIn1、In

2、In3、In4、In5と、出力ポートOut1、Out2、Out3、Out4、Out5を有する。ポートP1から入力したフレームは、シリアルパライズ変換装置SP1で、バッファメモリBM1と、8B10B変換コードDEC1を経由し、スイッチコントローラCTLと入力ポートIn1へ入力される。スイッチコントローラCTLにおいて、入力フレームのヘッダ部分に格納された送信先アドレスを解釈し、クロスバスイッチSWを切り換える。例として、ポートP2が出力先として選ばれた場合は、入力したフレームは出力ポートOut2と、8B10B変換エンコーダENC2と、バッファメモリBM2と、パライズシリアル変換装置PS2を経由して、ポートP2から出力される。ここで、バッファメモリBM1、BM2はFIFO (First-in First-out) メモリである。

[0016] シリアルパライズ変換装置SP1は、8B10B符号化されたシリアルデータを10ビット幅のパライズデータに変換し、ポートP1におけるデータ転送速度の1/10の速度に同期してバッファメモリBM1に書き込む。8B10BデコーダDEC1は、クロスバスイッチSWの動作速度に同期して、10ビットパライズデータをバッファメモリBM1から読み出し、8B10B復号化して、8ビットパライズデータに変換する。8B10BエンコーダENC2は、クロスバスイッチSWでスイッチされた8ビットパライズデータを再び8B10B符号化し、10ビットパライズデータに変換後、クロスバスイッチSWの動作速度に同期してバッファメモリBM2に書き込む。パライズシリアル変換装置PS2は、ポートP2におけるデータ転送速度の1/10の速度に同期して、10ビットパライズデータを10ビット幅のパライズデータに変換し、ポートP2から出力される。以上によりスイッチSW1は、ポートP1におけるデータ転送速度からポートP2におけるデータ転送速度へ速度変換する。

[0017] 図4は、ポートP1へ入力するフレームと、ポートP2、P3、P4、P5から出力されるフレームを示した図である。波形の凸はフレームが存在する時間、凹はフレームが存在していない時間を示している。フレームは伝送するデータ容量に従ってそのフレーム長が変化するが、ここではディスクアレイへのシーケンシャルアクセスが行われており、フレーム長が一定である。図4では、入力ポートP1でのデータ転送速度がある。図4では、入力ポートP1でのデータ転送速度が出力ポートP2、P3、P4、P5におけるデータ転送速度のM倍あるとする。従って、ポートP1におけるフレームM個の時間T1は、ポートP2から出力するフレームM個あるとする。従って、ポートP1におけるデータ転送速度は、ここでT3=M×T1である。

[0018] 入力データのデータ転送速度が速く、且つ出力データのデータ転送速度が遅い場合は、スイッチを周期的に切り換えないと出力ポートのバッファメモリが溢れ、スループットが低下する。フレームがスループットの低下無く

スイッチを通ずるには、図4のように周期的に出力ポートを切り換える必要がある。スイッチ切り替えポート数を n とすると、スイッチ切り替え周期 $T = n \times T1$ である（フレームの無い時間は無視した）。 $T2 \geq T3$ ならば、フレームの衝突無く、スループットの低下は起こらない。 $T2 \geq T3$ は $n \geq m$ と同じである。つまり、ディスクアレイへのデータ書き込み時に、スイッチにおいてスループット低下を起こさないための条件は、周期的に切り替えるスイッチポート数 n を、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度の、スイッチとディスクアレイ間のチャネル当りデータ転送速度に対する比 m 、以上に設定することである。この条件が保たれるれば、スイッチSW1は、ポートP1から入力したデータをバッファメモリにおいて速度変換し、フレーム単位で周期的に切り替えることにより逆多重化し、フレーム2、P3、P4、P5へ分配して出力する。スイッチを、周期的に切り換える方法の一つは、スイッチに接続したディスクグループをRAIDグループとすることである。RAIDのストライピング制に従えば、スイッチは周期的に切り替わる。

[0019] 図3は、ポートP2、P3、P4、P5からフレームを入力し、ポートP1から出力する場合を示す。これはディスクアレイからの読み出し時のスイッチ動作に相当する。例えば、ポートP2から入力したフレームは、シリアルラベル変換装置SP2と、バッファメモリBM2と、8B10B変換装置CODEC2を経由し、スイッチコントローラCTLと入力ポートin2へ入力される。スイッチコントローラCTLにおいて、入力フレームのヘッダ部分に書かれた送信先アドレスを解読し、クロススイッチXSWを切り換える。図3の場合は、ラウンドロビン方式によりクロスバススイッチXSWを切り替えて、順番にポートP2、P3、P4、P5から入力されるデータは全てポートP1へ出力する。すなわち、読み出し時は、複数の入力ポート（P2、P3、P4、P5）に同時にフレームが届く、これら複数の入力フレームは同期して入力ポートに届く必要はない。スイッチは、総当りの入力ポート間接続を切り替えることにより、これら複数の入力フレームを1フレームずつ出力ポート（P1）へ転送する。このように、スイッチを総当りの切り替える方式を、ラウンドロビン(Round Robin)方式と呼ぶ。ラウンドロビン方式により、結果的にスイッチは周期的に切り替わることになる。なお、読み出し時においても、スイッチはフレーム内送信優先制に従って切り替わることに違はない。フレームは出力ポートout1と、8B10B変換エンコードCODEC1と、バッファメモリBM1と、ラベルシリアル変換装置PS1を經由して、ポートP1から出力される。

[0020] シリアルラベル変換装置SP2は、8B10B符号化されたシリアルデータを10bit幅のバ

ラベルデータに変換し、ポートP2におけるデータ転送速度の $1/10$ の速度に同期してバッファメモリBM2に書き込む。8B10BデコードDEC2は、クロスバススイッチXSWの動作速度に同期して、10bitバレルデータデータをバッファメモリBM2から読み出し、8B10B復号化して、8bitバレルデータに変換する。8B10BエンコードENC1は、クロスバススイッチXSWでスイッチされた8bitバレルデータを再び8B10B符号化し、10bitバレルデータに変換し、ポートP1におけるデータ転送速度の $1/10$ の速度に同期して、10bitバレルデータをバッファメモリBM1に書き込む。ラベルシリアル変換装置PS1は、ポートP1におけるデータ転送速度の $1/10$ の速度に同期して、10bitバレルデータをバッファメモリBM1から読み出し、シリアル化して、ポートP1から出力する。以上よりスイッチSW1は、ポートP2におけるデータ転送速度からポートP1におけるデータ転送速度へ速度変換する。

[0021] 図5は、ポートP2、P3、P4、P5へ入力するフレームと、ポートP1から出力されるフレームを示した図である。波形の凸はフレームが存在する時間、凹はフレームが存在しない時間を示している。フレームは伝送するデータ容量に従ってそのフレーム長が変化するが、ここではディスクアレイへのシーケンシャルアクセスが行われており、フレーム長が一定である。図5では、入力ポートP1でのデータ転送速度が出力ポートP2、P3、P4、P5におけるデータ転送速度の m 倍であると、従って、ポートP5におけるフレームFe5の時間 $T4$ は、ポートP1からの出力時に $T5$ へ縮んでいる。ここで $T4 = m \times T5$ である。フレームFe2、Fe3、Fe4、Fe5をポートP1から出力するのにかかる時間を $T6$ とする。スイッチ切り替えポート数を n とすると、 $T6 \leq n \times T5$ である（フレームの無い時間は無視した）。スイッチにおいてフレームのスループット低下を防止するために、 $T6 \leq T4$ とする必要がある。 $T6 \leq T4$ は $n \leq m$ と同じである。

[0022] つまり、ディスクアレイからのデータ読み出し時に、スイッチにおいてスループット低下を起こさないための条件は、周期的に切り替えるスイッチポート数 n を、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度の、スイッチとディスクアレイ間のチャネル当りデータ転送速度に対する比 m 、以下に設定することである。この条件が保たれるれば、スイッチSW1は、ポートP2、P3、P4、P5から入力したデータをバッファメモリにおいて速度変換し、フレーム単位で周期的に切り替えることにより逆多重化し、フレーム2へ出力する。よって、ディスクアレイへの書き込みおよびディスクアレイからの読み出しを高速スループット化するためには、 $n \leq m$ 、つまり、周期的に切り替えるポート数を、ディスクアダプタとスイッチ間のチャネル当りデータ転送速度の、スイッチとディスクアレイ間のチャ

ネル当りデータ転送速度に対する比、と同程度に設定すればよいことが分かる。

[0023] 例えば、ディスクアダプタとスイッチ間の4Gbpsのチャネル1本で接続し、スイッチとディスクアレイ間を1Gbpsのチャネル4本で接続する。また、ディスクアダプタとスイッチ間の10Gbpsのチャネル1本で接続し、スイッチとディスクアレイ間を2Gbpsのチャネル4本で接続する。この場合、スイッチ出力ポート間でスループットのバランスが取れないので、有効的なスループットは $2Gbps \times 4 = 8Gbps$ となる。

[0024] 以上より、スイッチSW1において速度変換と多重化、逆多重化が行われるので、チャネルD1、D12、D13、D14上のデータ転送速度が低下し、チャネルD01、D02、D03、D04でのデータ転送速度は高速度になる。つまり、ディスクアダプタDKAとディスクアレイDA間のスループットを向上できる。本実施の形態のディスク装置におけるデータ転送方式としては、ファイバチャネルやインフィニバンドが使用できる。

[0025] 図6は、第1の実施の形態のディスク装置において、ディスクドライブの増設方法を示した図である。図6では図1に対して、ディスクグループR5とR6が増設されている。ディスクドライブを増設するた、スイッチSW1とSW2としてポート数の多いスイッチを使用して、ディスクドライブを増設すると、スイッチのディスクアレイ側のスループットが増加し、ディスクアダプタ側のスループットバランスが崩れるので、スイッチの速度変換機能が無効になる可能性がある。そこでスイッチSW1では、ディスクアダプタDKAとの間に、新規チャネルD05を増設している。また、スイッチSW2の場合は新規チャネルを増設せず、チャネルD02の信号伝送速度を増加させることで、ディスクアダプタ側とディスクアレイ側のスループットバランスを取っている。例えばスイッチSW1では、スイッチとディスクアレイ間を1Gbpsのチャネル4本で接続し、ディスクアダプタとスイッチ間を4Gbpsのチャネル2本で接続する。スイッチSW2では、スイッチとディスクアレイ間を1Gbpsのチャネル8本で接続し、ディスクアダプタとスイッチ間を10Gbpsのチャネル1本で接続する。このように、本実施の形態のディスク装置は、スイッチのポート数に応じて、ディスクドライブを増設可能である。このディスクドライブ増設方法は、1ポート当たり接続できるドラッグ数の少ないATA (AT Attachment) 方式ディスクドライブを増設するのに適用できる。

[0026] 図7に本発明の、第2の実施の形態であるディスク装置の構成を示す。本実施の形態のディスク装置は、第1の実施の形態のディスク装置に対して、ディスクアレイ部分の構成方法が異なる。本実施の形態のデ

ィスク装置は、ディスクコントローラDKCと、4個のディスクアレイDA1、DA2、DA3、DA4からなる。ディスクコントローラDKCは、チャネルアダプタCHA、キャッシュメモリCM、ディスクアダプタDKAからなる。ディスクアレイDA1とディスクアダプタDKAは、チャネルD01とスイッチSW1を介して接続する。同様に、ディスクアレイDA2はチャネルD02とスイッチSW2を介して、ディスクアレイDA3はチャネルD03とスイッチSW3を介して、ディスクアレイDA4はチャネルD04とスイッチSW4を介して、それぞれディスクアダプタDKAと接続する。スイッチSW1、SW2、SW3とSW4は、第1の実施の形態と同様に速度変換と多重化、逆多重化を行うスイッチとして機能する。本実施の形態におけるディスクアダプタDKAと、スイッチSW1、SW2、SW3、SW4と、ディスクアレイDA1、DA2、DA3、DA4との間のデータ転送方式は、ファイバチャネルを使用している。スイッチSW1、SW2、SW3、SW4はファイバチャネルスイッチである。

[0027] 本実施の形態におけるディスクアレイの構成を、ディスクアレイDA1を例に述べる。ディスクアレイDA1、DA2、DA3、DA4は、同様のドライブ構成である。ディスクアレイDA1は、チャネルD11上に接続した4個のディスクからなるディスクアレイと、D12上に接続した4個のディスクからなるディスクアレイと、D13上に接続した4個のディスクからなるディスクアレイと、D14上に接続した4個のディスクからなるディスクアレイ、からなる。チャネルD11を例にとると、ディスクドライブDK1、DK2、DK3、DK4が、チャネルD11上に接続されている。このように、多数のドライブを一つのチャネル上に接続してディスクドライブにアクセスする方法としては、ファイバチャネルアービトラリデッドループ（以下FC-A Lと呼ぶ）がある。

[0028] 図10に、FC-A Lの接続形態をディスクドライブDK1、DK2、DK3、DK4の接続形態を例として示す。各ディスクドライブの入出力ポートおよびスイッチSW1の入出力ポートは、送信機Txと受信機Rxを有する。FC-A Lの接続形態は、例えば図10に示すように、各ドライブの入出力ポートおよびスイッチの入出力ポートをループ状に接続するトポロジである。各ドライブの入出力ポートはファイバチャネルNL (Node Loop) ポートとして機能する。NLポートとは、ループ動作をする装置（ここではディスクドライブ）のポートである。スイッチSW1のディスクアレイDA1接続側出力ポートは、ファイバチャネルのFL (Fabric Loop) ポートとして機能する。FLポートとは、FC-A Lを接続可能なスイッチのポートである。FLポートを有するループは、ファイバチャネルのパブリックループとして機能するので、

チャネルD11が形成するFC-A1はパブリックグループとなる。パブリックグループとは、ループ上のディスクドライブが、スイッチを介してループ外のポートと通信可能なループである。よって、ディスクドライブDK1、DK2、DK3、DK4は、スイッチSW1およびチャネルD01を介してディスクアダプタDKAと通信可能である。以上、チャネルD13の接続形態を例に説明したが、チャネルD12、D13、D14でも同様である。本実施の形態のディスク装置においてRAIDシステムを構築する場合は、ディスクグループR1、R2、R3、R4を、それぞれRAIDグループとする。本実施の形態では、4個のディスクドライブでRAIDグループを構成しているが、RAIDグループを構成するドライブ数を4個に限るものではない。

05

チャネルD001を介してディスクアダプタDKAと通信可能である。以上、チャネルD11の接続形態を例に説明したが、チャネルD12、D13、D14でも同様である。本実施の形態のディスク装置においてRAIDシステムを構築する場合は、ディスクグループR1、R2、R3、R4を、それぞれRAIDグループとする。本実施の形態では、4個のディスクグループでRAIDグループを構成しているが、RAIDグループを構成するドライブ数を4個に限るものではない。

【0029】本実施の形態においては、チャネルD1、D12、D13、D14において、それぞれFC-ALを用いてディスクドライブを接続している。FC-

ALの仕様から、チャネルD11、D12、D13、D14上には、それぞれ最大126台までのディスプレイが接続可能である。また、チャネルD01、D02、D03、D04の媒体として光ファイバケーブルを、チャネルD11、D12、D13、D14の媒体としてメタルケーブルを用いる。

【0030】以上説明したように、本実施の形態のディスプレイ装置においては、ディスプレイドライバをFC-A1で接続しているため、スイッチのポート当りに接続できるドライバ台数が増加する。また、ディスプレイ当りの記憶容量を増加させる効果がある。また、ディスプレイをメタルケーブルで接続することにより、ディスプレイ毎に高画質光インターフェイスを装備する必要があるが、ディスプレイのコストを下げる効果がある。

【0031】図8に本発明の、第3の実施の形態であるディスク装置の構成を示す。本実施の形態のディスク装置は、ディスクコントローラとスイッチを二重化した点に特徴がある。本実施の形態において、ディスクアダプタDKA1、DKA2と、スイッチSW1、SW2と、ディスクアレイDA1ととの間のデータ転送方式は、ファイバチャネルを使用している。本実施の形態のディスク装置は、ディスクコントローラDKC1、DKC2と、スイッチSW1、SW2と、ディスクアレイDA1からなる。スイッチSW1とSW2は、第1の実施の形態と同様に速度変換と多重化、逆多重化を行うスイッチとして機能する。ディスクコントローラDKC1は、チャネルアダプタCHA1と、キャッシュメモリCM1と、ディスクアダプタDKA1からなる。ディスクコントローラDKC2は、チャネルアダプタCHA2と、キャッシュメモリCM2と、ディスクアダプタDKA2からなる。ディスクアダプタDKA1とスイッチSW1をチャネルD1aで接続し、ディスクアダプタDKA2とスイ

107

ツチSW2をチャネルD2aで接続し、ディスプレイアダプタDKA1とスイッチSW2をチャネルD1bで接続し、ディスプレイアダプタDKA2とスイッチSW1をチャネルD2bで接続する。

【0032】ディスクアレイDA1を構成するディスクドライブは、出入口ポートを2個有する。例えば、ディスクドライブDK1、DK2、DK3、DK4は、チャネルD11およびD21の両チャネルと接続する。ディスクアレイDA1は、チャネルD11とD21に接続した4個のディスクからなるディスクアレイと、D12とD22に接続した4個のディスクからなるディスクアレイと、D13とD23に接続した4個のディスクからなるディスクアレイと、D14とD24に接続した4個のディスクからなるディスクアレイ、からなる。チャネルD11、D12、D13、D14、D21、D22、D23、D24は、FC-AIでディスクドライブを接続する。

【0033】図11に本実施の形態におけるFCC-A1の接続形態を、ディスクドライブDK1、DK2、DK3、DK4の接続形態を例として示す。各ディスクドライブは、それぞれNLポートを有する。各ディスクドライブの入出力ポートおよびスピンチSW1、SW2の入出力ポートは、送信機Txと受信機Rxとを有する。スピンチSW1、SW2のディスクアレイDA1接続側

入出力ポートは、FLポートである。チャネルD11により、スイッチSW1、ディスプレイドライブDK1、DK2、DK3、DK4をループ状に接続する。同様にチャネルD21により、スイッチSW2、ディスプレイドライブDK1、DK2、DK3、DK4をループ状に接続する。これら2個のループは、ファイバチャネルのパブリックグループであり、ディスプレイドライブDK1、DK2、DK3、DK4は、スイッチSW1またはSW2を介してディスプレイドライブDKA1またはDKA2と通信可能である。以上、チャネルD11、D21の接続形態を例に説明したが、チャネルD12、D13、D14、D22、D23、D24でも同様である。本実施の形態のディスプレイ装置においてRAIDシステムを構成する場合、ディスプレイグループR1、R2、R3、R4を、それぞれRAIDグループとする。本実施の形態では、4個のディスプレイでRAIDグループを構成しているものが、RAIDグループを構成するドライブ数を4個に限るものではない。

【0034】 ディスクアレイDA1内の全ディスクドライ
イブは、ディスクアダプタDKA1およびDKA2のど
ちらからでもアクセス可能である。本実施の形態のディ
スク装置は、チャネルD1b、D2bをスイッチSW
1、SW2故障時の迂回経路として使用する。例えばス
イッチSW1が故障した場合でも、ディスクアダプタ
DKA1はチャネルD1bとスイッチSW2経由でディ
スクアレイDA1にアクセスできる。逆に、ス
イッチSW2が故障した場合でも、ディスクアダプタDKA2はチャネルD2bとスイッチSW1経由でディ
スクアレイDA1にアクセスできる。

. 2003 11 18 13:49

2が故障した場合は、ディスクアダプタDKA2はチャネルD2bとスイッチSW1経由でディスクアレイDA1にアクセスできるので、信頼性が高いディスク装置が実現できる。

【0036】図12に本発明の、第4の実施の形態であるディस्क装置の構成を示す。本実施の形態のディスク装置は、第3の実施の形態のディスク装置に対して、スイッチSW1、SW2間を接続するチャネルD3 a、D3 bを設けた点に特徴がある。本実施の形態において、ディスクアダプタDKA1、DKA2と、スイッチSW1SW2 1、SW2と、ディスクアレイDA1と、間のデータ転送方式は、ファイバチャネルを使用している。本実施の形態のディスク装置は、ディスクコントローラDKC1、DKC2と、スイッチSW1、SW2と、ディスクアレイDA1からなる。スイッチSW1とSW2は、第1の実施の形態と同様に速度増進と多重化、逆多重化を行うスイッチとして機能する。ディスクコントローラDKC1は、チャネルアダプタCHA1と、キャッシュメモリCM1と、ディスクアダプタDKA1からなる。ディスクコントローラDKC2は、チャネルアダプタCHA2と、キャッシュメモリCM2と、ディスクアダプタDKA2からなる。ディスクアダプタDKA1とスイッチSW1をチャネルD1 aで接続し、ディスクアダプタDKA2とスイッチSW2をチャネルD2 aで接続し、ディスクアダプタDKA1とスイッチSW2をチャネルD1 bで接続し、ディスクアダプタDKA2とスイッチSW1をチャネルD2 bで接続する。さらに、スイッチSW1とSW2をチャネルD3 a、D3 bで接続する。【0036】ディスクアレイDA1を構成するディスクドライブは、入出力ポート2個有する。例えば、ディスクドライブDK1、DK2、DK3、DK4は、チャネルD1およびD2の1両チャネルと接続する。ディスクアレイDA1は、チャネルD11とD211に接続した4個のディスクからなるディスクアレイと、D12とD22に接続した4個のディスクからなるディスクアレイと、D13とD23に接続した4個のディスクからなるディスクアレイと、D14とD24に接続した4個のディスクからなるディスクアレイ、からなる。チャネルD11、D12、D13、D14、D21、D22、D23、D24は、図11に示すようにFC-ALでディスクドライブを接続する。ディスクアレイDA1内の全ディスクドライブは、ディスクアダプタDKA1およびDKA2のどちらからでもアクセス可能である。本実施の形態のディスク装置においてRAIDシステムを構築する場合は、ディスクグループR1、R2、R3、R4を、それぞれRAIDグループでRAIDシステムを構成しているが、RAIDグループを構成するドライブ数を4個に限るものではない。

【0037】ディスクアダプタDKA1、DKA2とマ

- 10 -

2003 11 18 13:49

sとなる。第3の実施の形態において、ディスクアダプタ-ディスクアレイ間スループットを4Gbpsにするためには、チャネルD1aおよびD2aのデータ伝送速度を、それぞれ4Gbpsに高める必要がある。以上から、本実施の形態によれば、ディスクアダプタ-スイッチ間のチャネル当りデータ伝送速度が低くても、ディスクアダプタ-ディスクアレイ間の総スループットが高いディスク装置が実現できる。

【0039】

【発明の効果】以上説明したように、本発明によれば以下の効果がある。ディスクアダプタとディスクアレイ間のスループットが高いディスク装置を提供できる。また、ディスクアダプタとディスクアレイ間のスループットが高く、且つディスクドライブ接続台数が多いディスク装置を提供できる。また、信頼性の高いディスクアレイを有するディスク装置を提供できる。また、信頼性が高いディスクアダプタとディスクアレイ間ネットワークを有するディスク装置を提供できる。また、信頼性およびスループットが高いディスクアダプタとディスクアレイ間ネットワークを有するディスク装置を提供できる。また、ディスクからの読み出しおよびディスクへの書き込みを高スループット化できるディスク装置を提供できる。また、高スループットを維持できるディスク装置を提供できる。また、ディスクアダプタとディスクアレイ間のスループットが高く低コストなディスク装置を提供できる。

【図面の簡単な説明】

【図1】本発明の第1の実施の形態のディスク装置を示す図である。
【図2】本発明に用いるスイッチの構成を示す図である。
【図3】本発明に用いるスイッチの構成を示す図である。
【図4】本発明に用いるスイッチの動作を示す図である。
【図5】本発明に用いるスイッチの動作を示す図である。
【図6】本発明第1の実施の形態に対して、ディスクドライブを増設する方法を示す図である。

【図7】本発明の第2の実施の形態のディスク装置を示す図である。

【図8】本発明の第3の実施の形態のディスク装置を示す図である。

05 【図9】従来のディスク装置を示す図である。

【図10】FC-ALEによる接続形態を説明する図である。

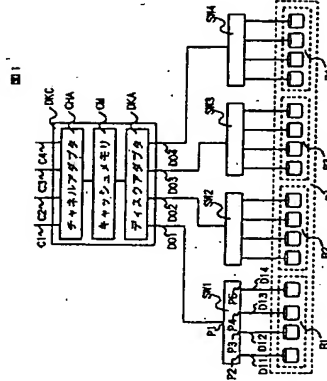
【図11】FC-ALEによる接続形態を説明する図である。

10 【図12】本発明の第4の実施の形態のディスク装置を示す図である。

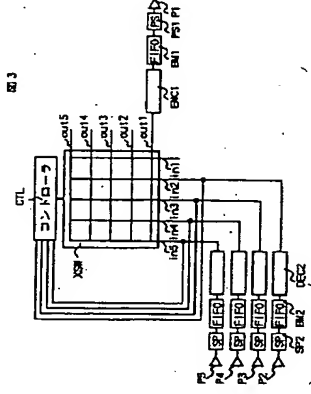
【符号の説明】

DKC、DKC1、DKC2 ディスクコントローラ
CHA、CHA1、CHA2 チャネルアダプタ
15 CM、CM1、CM2 キャッシュメモリ
DKA、DKA1、DKA2 ディスクアダプタ
DA、DA1~DA4 ディスクアレイ
DK1~DK4 ディスクドライブ
R1~R6 ディスクグループ
20 C1~C4、D1~D4、D01~D05、D11~D14、D21~D24、D1a、D1b、D2a、D2b、D3a、D3b チャネル
SW1~SW4 スイッチ
P1~P5 スイッチポート
25 XSW クロスバススイッチ
CTL スイッチコントローラ
in1~in5 クロスバススイッチ入力ポート
out1~out5 クロスバススイッチ出力ポート
SP1、SP2 シリアルパラレル変換装置
30 PS1、PS2 パラレルシリアル変換装置
BM1、BM2 バッファメモリ
DEC1、DEC2 8B10B変換デコーダ
ENC1、ENC2 8B10B変換エンコーダ
T1、T2、T3、
35 T4、T5、T6 フレームの時間
Tx 送信機
Rx 受信機
NL NLポート
FL FLポート

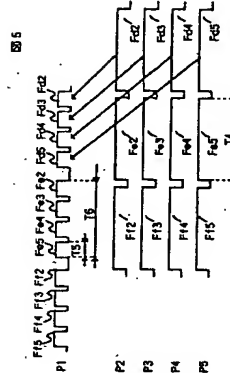
【図1】



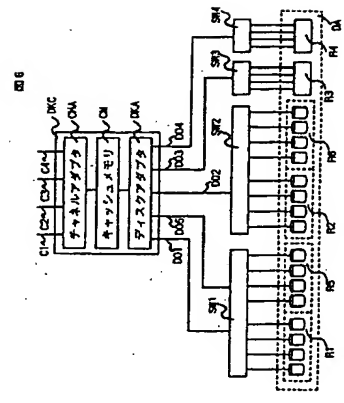
【図3】



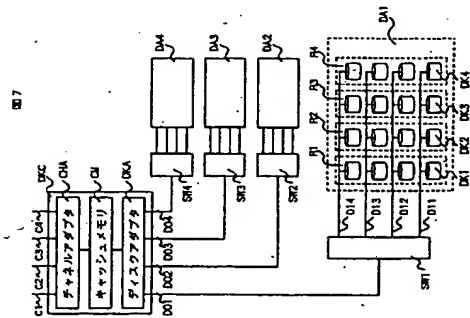
【図5】



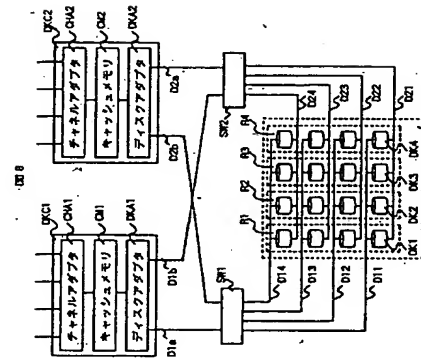
【図6】



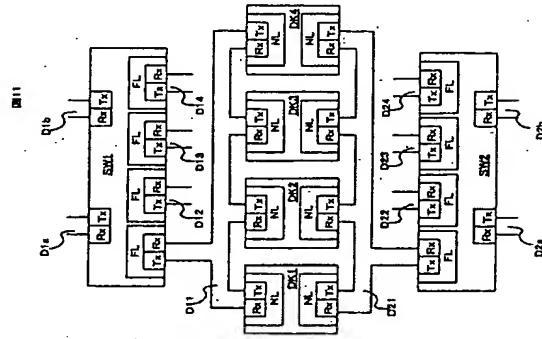
【圖7】



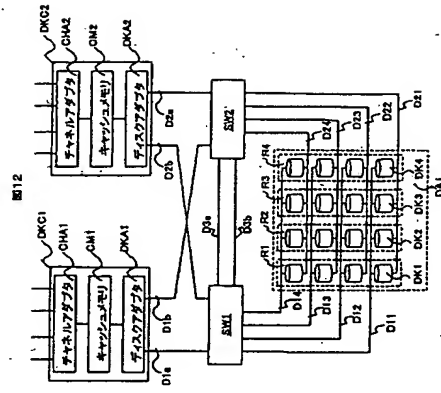
【88】



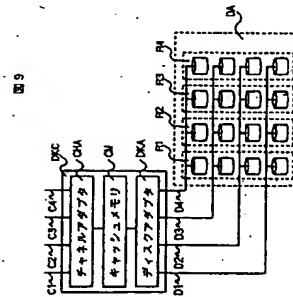
【图 1-1】



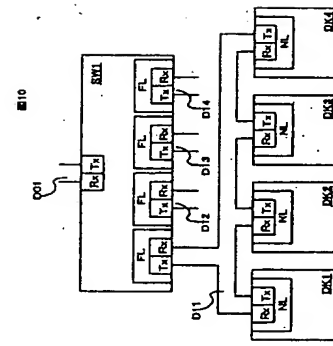
[12]



【6】



【图10】





Simplify™

**Network Storage Group
Host Products
Technology Brief**
October 8, 2002

FULL-DUPLEX AND FIBRE CHANNEL

WHAT IS FULL-DUPLEX?

With full-duplex data transmission, data is received and transmitted at the same time. A Fibre Channel adapter that has full-duplex capability can send data to a Fibre Channel node and receive data from that node simultaneously.

WHY FIBRE CHANNEL?

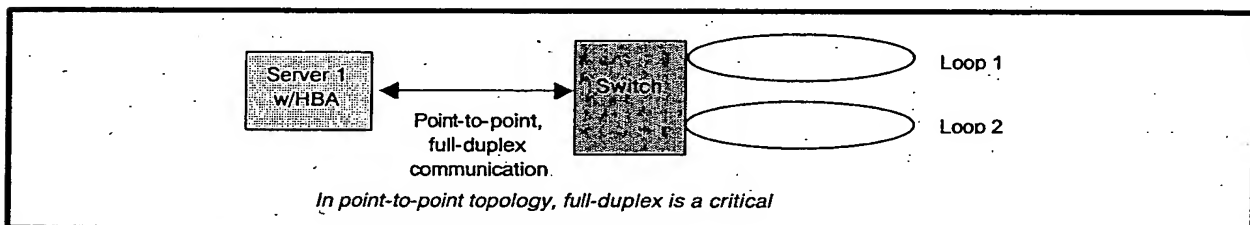
Traditional I/O channels are buses. Buses are like one-way tunnels. Parallel SCSI and ATA, for example, can only process a single point-to-point transfer at a time. Fibre Channel is different. On a single Fibre Channel cable, there are two connections between any two devices:

- The outbound half of the cable goes from the transmitting device to the receiving device.
- The inbound half of the cable goes from the receiver back to the sender and completes the connection.

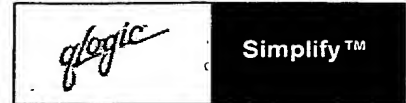
These connections are physically separate; Fibre Channel supports separate communications being in process on each half of the cable at the same time. This capability, called full-duplex communication, makes Fibre Channel more efficient than traditional buses.

WHEN IS FULL-DUPLEX MOST CRITICAL?

Full-duplex is most important in point-to-point communications on a switched fabric. A host system communicating with a switch can take advantage of the simultaneous send and receive capabilities. A switch is most likely to have data ready to transmit to the host when a connection is opened between them.



In the above diagram, the host adapter in the Server 1 is communicating simultaneously with a drive on Loop 1 and a drive on Loop 2. (The drives can be spread across more than two loops; however, only two loops are needed to illustrate the feature.) In this configuration, the system can realize the performance potential of Fibre Channel full duplex capability.



SUMMARY

Full-duplex is one of many features that highlight the advantages of Fibre Channel. When a host adapter accesses drives on multiple loops on a switch and has a workload that keeps the attached drives busy, using full-duplex results in a considerable improvement in performance.